UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE BIOLOGIA ANIMAL

Molecular pathogenesis of a malformation syndrome associated with a pericentric chromosome 2 inversion

Manuela Pinto Cardoso

Mestrado em Biologia Humana e do Ambiente

Dissertação orientada por:
Doutor Dezsö David
Doutora Deodália Dias

2017

ACKNOWLEDGEMENTS

I would like to say “thank you!” to all the people that contributed in some way to this thesis.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Dezsö David, for giving me the opportunity to work in his research group and for everything he taught me. Without his mentorship I would have never learned so much.

I am grateful for Prof. Deodália Dias’s encouragement and support in all these years that I have been under her wings.

I would like to extent my thanks to everyone at the National Health Institute Dr. Ricardo Jorge, for their continuous help in all stages of this thesis. To the team at Harvard Medical School, thank you for the technical assistance, and in special Dr. Cynthia Morton and Dr. Michael Talkowski.

I am also grateful to Dr. Rui Gonçalves and Dr. João Freixo, who accompanied this case study and shared their medical knowledge. Of course, I am grateful for the family members for their involvement in this study.

To my lab mates, a shout-out to them all! I really hold them dear for their help and the many laughs we shared every day. Thank you Mariana for being there literally since day one and for playing the role of a more mature counterpart. To Joana for being a glitter of fun and for reminding me of my curiosity in all things techy and kind of odd. To Sara for your positiveness. And Raquel, Carlos and Inês, though now so far away, it was a pleasure crossing paths with you.

To my friends, both online and offline, for your words of encouragement. To Mapril, without whom I would be so very lost.

At last, but obviously not the least, my dearest family, specially my very patient parents. They who silently watched over me and gave me the strength to face any adversity. They who so unwearyingly accepted my stubbornness. They are the guiding light in the darkest of hours. I do not say this often enough: I love you and thank you so much for being with me. Always.

ABSTRACT

Congenital malformation syndromes can be caused by genomic and/or chromosome rearrangements. It is difficult to establish the underlying causes of malformations because of their high level of complexity. Although balanced chromosome inversions are in most cases subclinical, those disrupting transcripts or affecting the genomic architecture at breakpoint regions may well be pathogenic. Currently, the lack of a fully annotated human genome hinders the predictability of the phonotypic consequences of such rearrangements.

The aim of this study is the identification of potential candidate genes for a malformation syndrome in an individual with an apparently balanced maternally inherited pericentric chromosome inversion inv(2)(p16.1;q14.3)mat. The proband has severe congenital malformation with multiple psychomotor and developmental anomalies, dismorphism and autistic features. The parents are phenotypically normal.

Classical cytogenetic methods are of low resolution, often in the magnitude of a 5 to 10 Mb. Whole-genome Next-Generation Sequencing (NGS) of large-insert sequencing library (liWGS) has the capability to detect structural rearrangements with incomparably higher resolution, including cryptic alterations. As consequence, it was applied for the identification of inv(2)(p16.1;q14.3) breakpoints in the proband. Familial segregation analysis and definition of the inversion breakpoints at a nucleotide resolution were performed by amplification of junction fragments and Sanger sequencing. Genome and transcriptome array analysis were also carried out, for detection of additional genomic alterations and for gene expression profiling, respectively.

Additionally, a possibly polymorphic duplication at 2q21.1, inherited from his father, was found. No apparent pathogenic genomic imbalances were identified in the proband.

The inversion breakpoints are located at chr2:55,935,064 and chr2:123,767,685 (GRCh37), respectively, in 2p16.1 and 2q14.3. The inv2p16.1 breakpoint is flanked 14 kb proximal by the gene polyribonucleotide nucleotidyltransferase 1 (*PNPT1*; chr2:55,861,198-55,921,045, GRCh37; OMIM *610316) and 172 kb distal by EGF containing fibulin-like extracellular matrix protein 1 (*EFEMP1*; chr2:56,093,097-56,151,298, GRCh37; OMIM *601548). *PNPT1*, highly expressed in mice cochlea, has been associated with deafness (OMIM #614934) and with combined oxidative phosphorylation deficiency (OMIM #614932), both autosomal recessive. Meanwhile, the autosomal dominant Doyne honeycomb retinal dystrophy (OMIM #126600) is reported to be associated with mutations in *EFEMP1*. This gene is essential for the formation of elastic fibers in connective tissue.

The 2q14.3 breakpoint is in a gene-poor region. Located 1.2 Mb proximal to the breakpoint is translin (*TSN*; chr2:122,513,120-122,525,428, GRCh37; OMIM *600575). Involved in DNA damage repair and RNA trafficking in neurons, *TSN* codes for a protein that specifically binds to breakpoint junctions of translocations in acute leukemia. The gene contactin-associated protein-like 5 (*CNTNAP5*; chr2:124,782,863-125,672,953, GRCh37; OMIM *610519) is localized 1 Mb, distal. *CNTNAP5* is involved in cell adhesion and intercellular communication. Susceptibility to autistic syndromes has been suspected.

The above described breakpoints at nucleotide resolution are the same in the proband's mother, and did not directly disrupt any gene.

Publicly available clinical information on alterations affecting the inversion flanking genes revealed no major similarity with the proband's phenotype. Furthermore, no significant alteration in their expression level was observed. In-depth analysis of genome-wide expression data is in progress.

Based on these findings, the causal relationship between clinical phenotype and the inv(2)(p16.1;q14.3) is most likely excluded, since the inversion is most likely non-pathogenic. Therefore it is not yet possible to identity the underlying genetic cause of the malformation syndrome reported in this subject. Whole-exome sequencing is proposed as a future task to detect the disease causing alteration.

This study highlights the application of NGS-based methodology, with its capability in mapping chromosome inversion breakpoints at a very high resolution. Large scale application of this approach will represent a hallmark in the characterization of congenital malformations associated with structural chromosomal abnormalities.

This study was supported by *Fundação para a Ciência e a Tecnologia* project PTDC/SAU-GMG/118140/2010 and HMSP-ICT/0016/2013

Keywords
Congenital malformation syndrome; Chromosome 2 inversion; NGS technology; *PNPT1*; *CNTNAP5*

RESUMO

As síndromes de malformação congénitas são um dos principais grupos de patologias que afetam neonatos e crianças em países desenvolvidos. Muitos destes casos têm como base genética os arranjos genómicos ou cromossómicos. No entanto, por norma, devido à complexidade inerente às síndromes de malformação, é difícil e laborioso identificar com exatidão a alteração molecular que lhes deu origem. Aliado à inexistência atual de um genoma humano completamente anotado, torna-se complicado a compreensão e a previsão das consequências fenotípicas dos rearranjos cromossómicos.

As inversões cromossómicas são rearranjos que ocorrem quando dois pontos de quebra ocorrem num mesmo cromossoma e são reinseridos invertidos, sem alteração de número de cópia. Normalmente as inversões são subclínicas, sem um fenótipo clínico associado. Se estes forem transmitidos a mais de 1% de uma dada população, tratam-se de polimorfismos. Se um rearranjo afectar transcritos ou a arquitetura genética junto dos pontos de quebra, perturbando assim o normal funcionamento dos genes, sobretudo os de expressão indispensável, este estará envolvido na etiologia de uma patologia potencialmente grave. Comparado com outros rearranjos cromossómicos, são poucas as inversões atualmente detalhadamente caracterizadas, frequentemente devido a dificuldades técnicas relacionadas com regiões repetitivas, frequentes nos pontos de quebra das inversões.

Metodologias clássicas de citogenética são de baixa resolução e por vezes incapazes de identificar determinadas anomalias estruturais. As tecnologias atualmente mais avançadas para o estudo de rearranjos incluem *microarrays* genómicos, ideal na análise de variações no número de cópias, e a sequenciação de próxima geração (NGS), mais concretamente sequenciação pangenómica, para a generalidade dos rearranjos cromossómicos. Esta última tem a particularidade de ser eficiente na identificação de alterações crípticas, de oferecer uma potencial resolução bastante elevada (em certos casos nucleotídica) e de gerar grande quantidade de dados rapidamente. Das plataformas NGS existentes, as mais aptas para a análise de inversões envolvem a construção de bibliotecas *mate-pair* de grandes insertos, cuja distância entre pares de leitura é de 2 a 6 kb, permitindo superar dificuldades técnicas com zonas repetitivas e pequenas alterações junto aos pontos de quebra.

Esta tese pretende identificar as alterações moleculares responsáveis pela síndrome de malformação congénita num indivíduo portador de uma inversão cromossómica pericêntrica aparentemente equilibrada de origem materna.

O caso índex, portador da síndrome de malformação, apresenta acentuado atraso de desenvolvimento mental e psicomotor, dismorfia facial e perturbações do espectro do autismo. Ele tem muito baixo peso e altura para a idade. Foram também diagnosticadas cardiopatias, criptorquidia, escoliose e hipotonia generalizada. Estudos citogenéticos detetaram a existência de uma inversão pericêntrica no cromossoma 2, também encontrada na mãe. Os pais têm fenótipo aparentemente normal.

Primeiramente, procedeu-se à identificação de alterações estruturais desequilibradas no indivíduo índex. Foram detetaram várias alterações de número de cópia, na maioria pequenas (< 100kb) e sem envolver genes OMIM, com a exceção da duplicação de 590 kb em 2q21.1. Os genes na duplicação não aparentam estar relacionados com o fenótipo observado. Ademais, foi detetado uma duplicação de 610 kb no pai nesta mesma região genómica, sugerindo que se trata de uma alteração de origem paterna, muito provavelmente não-patogénica e possivelmente de natureza polimórfica.

Sequenciação pangenómica de grandes insertos (*large-insert whole-genome sequencing*) usando ácido desoxirribonucleico (ADN) do caso índex foi realizada para a identificação dos pontos de quebra da inversão no cromossoma 2. Uma vez delimitado a região dos pontos de quebra por NGS, foram desenhados oligonucleotídeos específicos para a amplificação dos fragmentos de junção e, seguidamente, procedeu-se à análise de segregação familiar e determinação nucleotídica dos pontos de quebra através de sequenciação Sanger. O estudo do perfil de expressão genética foi feito com *Human Transcriptome Assay* (HTA 2.0) da Affymetrix, utilizando ácido desoxirribonucleico (ARN) da linha celular linfoblastóide do indivíduo índex.

Os dados obtidos por NGS permitiram a redefinição da localização genómica da inversão. O cariótipo do caso índex foi assim redefinido como 46, XY, inv(2)(p16.1q14.3)mat.

Os pontos de quebra da inversão no cromossoma 2, no caso índex e na sua mãe, foram determinados. Estes localizam-se na posição chr2:55,935,064 e chr2:123,767,685 (GRCh37), respetivamente, nas bandas p16.1 e q14.3. Na sequência invertida ocorreu a deleção de 5 bases. Os pontos de quebra da inversão são iguais em ambos os indivíduos, sem quaisquer alterações detetadas nos fragmentos de junção. Segundo a nomenclatura baseada em citogenética de próxima geração, esta inversão é descrita como seq[GRCh37] inv(2)(pter→2p16.1(55,935,06{1-3})::2q14.3(123,767,68{3-1})→2p16.1 (55,935,06{5-4})::2q14.3(123,767,68{4-5})→qter).

Os pontos de quebra não interrompem diretamente genes conhecidos. Em inv2p16.1, este é flanqueado a 5' pelo gene polirribonucleotídeo nucleotidiltransferase 1 (*PNPT1*; chr2:55,861,198-55,921,045, GRCh37; OMIM *610316), e a 3' pelo gene proteína 1 da matriz extracelular tipo-fibulina contendo EGF (*EFEMP1*; chr2:56,093,097-56,151,298, GRCh37; OMIM *601548) a 158 kb. O *PNPT1* está envolvido na cadeia respiratória mitocondrial. Mutações em homozigotia foram associadas com deficiência na fosforilação oxidativa (OMIM #614932), originando nomeadamente encefalopatias, e com a surdez hereditária autossómica recessiva 70 (OMIM #614934). Em murganhos, tem expressão acentuada na cóclea. *EFEMP1* é essencial para a correta formação de fibras elásticas em tecido conjuntivo, tendo elevada expressão nos pulmões e esófago em murganhos, e baixa no cérebro e coração. Mutações neste gene estão descritas como causa genética da distrofia da retina de Doyne (OMIM #126600), patologia autossómica dominante com início na segunda década de vida, causando perda progressiva de visão.

O ponto de quebra em inv2q14.3 situa-se numa região pobre em genes. O gene translina (*TSN*; chr2:122,513,120-122,525,428, GRCh37; OMIM *600575) franqueia o ponto de quebra proximamente a 1240 kb, enquanto o gene tipo-proteína associada à contatina 5 (*CNTNAP5*; chr2:124,782,863-125,672,953, GRCh37; OMIM *610519) localiza-se 1020 kb distal do ponto de quebra. *TSN* codifica uma proteína que reconhece sequências-alvo em junções de pontos de quebra de translocações em doentes com leucemia, e está envolvido no mecanismo de reparação de ADN e transporte de ARN em neurónios. Em murganhos, expressa-se preferencialmente no tecido adiposo. O *CNTNAP5* produz uma proteína que atua no sistema nervoso como moléculas de adesão celular e de recetor na comunicação intercelular. Em murganhos, expressa-se predominantemente no sistema nervoso. Existe suspeita de que mutações pontuais possam conferir suscetibilidade a comportamentos do espectro do autismo.

Quanto à expressão genética, os resultados mostraram que os genes que flanqueiam a inversão não aparentam ter nível de expressão significativamente alterada comparativamente com os controlos. O estudo aprofundado de expressão a nível genómico está a decorrer.

Os restantes genes próximos dos pontos de quebra da inversão relevaram baixa probabilidade de serem as alterações causadoras do fenótipo, nomeadamente a nível das doenças associadas.

Tendo em conta os resultados obtidos, especialmente a confirmação da origem materna da inversão, esta alteração não aparenta ser a principal e única causa molecular do fenótipo. Ademais, esta conclusão é suportada pela pouca sobreposição clínica dos genes flanqueadores com a síndrome de malformação congénita, e da expressão génica aparentemente não alterada.

Assim, atualmente, a relação causal entre o fenótipo observado e a inversão no cromossoma 2 foi excluída. Esta inversão é muito provavelmente não-patogénica por si só. Até ao momento e com os dados disponíveis, não foi possível identificar genes candidatos nem as alterações moleculares por detrás da síndrome de malformação congénita no caso índex. Informação médica disponível exclui influência de fatores ambientais na embriogénese.

Futuramente, sugere-se recorrer à sequenciação do exoma, visto que tem uma sensibilidade muito superior para a deteção de pequenas em exões, potencialmente não detestáveis pelas abordagens até ao momento utilizadas. Adicionalmente, o estudo nos restantes membros da família permitirão obter uma melhor visão sobre a segregação familiar.

Este estudo teve o suporte do projeto da Fundação para a Ciência e a Tecnologia *HMSP-ICT/0016/2013*.

Palavras-chave
Síndrome de malformação congénita; Inversão cromossoma 2; Tecnologias NGS; *PNPT1*; *CNTNAP5*

INDEX

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Array CGH	Array comparative genomic hybridization
ASD	Autism spectrum disorder
Bp	Base pair
cDNA	Complementary DNA
CNV	Copy number variation
cRNA	Complementary RNA
DGV	Database of Genomic Variants
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
EDTA	Ethylenediaminetetraacetic acid
FBS-Hi	Heat inactivated fetal bovine serum
GEO	Gene Expression Omnibus
GRCh37	Genome Reference Consortium, build 37
HTA 2.0	Human Transcriptome Array 2.0
HuGene 1.0 ST	Human Gene 1.0 ST Array
Kb	Kilobase pair
LCL	Lymphoblastoid cell lines
LINE	Long interspersed nuclear elements
liWGS	Large-insert whole-genome sequencing
Mb	Megabase pair
MIM	Mendelian Inheritance in Man
NCBI	National Center for Biotechnology Information
NGS	Next-generation sequencing
OMIM	Online Mendelian Inheritance in Man
PCR	Polymerase chain reaction
RNA	Ribonucleic acid
SD	Segmental duplication
SINE	Short interspersed nuclear elements
SNP	Single-nucleotide polymorphism

1. INTRODUCTION

1.1. The Chromosomes

Since ancient times that humans inquired about the inheritance of traits from generation to generation. It was not until 1860 that the existence of a biological element of heredity was proposed. In 1869, the deoxyribonucleic acid (DNA) fiber was discovered, with its role in heredity finally confirmed in 1952 and its structure famously uncovered a year later (Bhat and Wani 2017; Dahm 2007; Watson and Crick 1953).

The long chain of DNA is associated with proteins called histones that compact it, protect it and help in the correct function of the DNA. The DNA and the associated components are packaged into thread-like structures, the chromatin, which in turn make up the chromosomes (Paulson and Vagnarelli 2011). The word chromosome derives from the Greek words for "color" and "body", due to their strong coloration when stained with specific dyes. Chromosomes are the structures that hold most of an organism' genes, the individual instructions for development and function, governing every characteristic of an organism (GHR 2014). Over time, the genetic information coded in the DNA is transmitted from parent cells to daughter cells, as well as from parent to child (Czepulkowski 2001).

Each chromosome has a primary constriction called the centromere, dividing it into two sections, known as arms, the short p arm and the long q arm. At the extremity of each arm is a region called telomere (GHR 2014; Young 2005). A chromosome can be classified as metacentric, submetacentric or acrocentric depending on whether the centromere is central, slightly off-center or almost at the telomere, respectively (Czepulkowski 2001).

The number of chromosomes varies between species and, within an organism, whether it is a gamete or a somatic cell, being respectively haploid and diploid for human (Czepulkowski 2001). The normal human chromosome complement consist of 46 chromosomes (23 pairs), divided into 44 autosomes and a pair of sex chromosomes (Young 2005). Females have two copies of the X chromosome, while males have one X and one Y chromosome. The 22 autosome pairs are numbered by size. When lined up in pairs from the largest to the smallest chromosome, with the short arms at top, it is called a karyotype (GHR 2014).

1.2. Chromosomal Anomalies

Chromosomal anomalies traditionally refer to alterations in the chromosome that are generally large enough to be visible under a light microscope, which has an average resolution of 5 to 10 Mb (Czepulkowski 2001). With the development of much more sensitive technologies, gradually the criteria for the designation of structural anomalies have been shifting to a magnitude of thousands of bases (Vergult et al. 2014).

These anomalies are classified as either numerical or structural and may involve more than one chromosome. Numerical anomalies are deviations from the normal state of diploidy. In another hand, structural anomalies, also called chromosomal rearrangements, changes the very structure of one or

more chromosomes, such as the case of translocation, deletions or inversions (Young 2005). The anomalies may cause profound alterations that in turn may disrupt normal cell function, if such events could not be rapidly mended correctly by the DNA repairing mechanisms (Czepulkowski 2001).

1.2.1. Structural Anomalies

A structural rearrangement encompasses several events that together alter the chromosome morphology. These events are caused by double strand breaks in the DNA, usually at two different loci, followed by a rejoining of the broken ends to produce a new arrangement with the corresponding consequences (Griffiths et al. 1999). When these rearrangements involve more than two chromosomes or breakpoints they are termed complex chromosome rearrangements (Liu et al. 2011).

A variety of cellular processes can act as the source that give rise to DNA breaks. During DNA replication, for example, if a replication fork passes through a template that contains a single-stranded break, it will be converted into a double-stranded break on one of the sister chromatid thus creating a full breakpoint (van Gent et al. 2001). External factors such as ionizing radiation are also capable of inducing breaks. Several mechanisms are available to repair these damages, but they can sometimes lead to other small errors upon conclusion: homologous recombination repair precisely restores the original sequence at the break without modifications; single-strand annealing could lead to interstitial deletions; and, finally, nonhomologous DNA end joining connects two broken ends directly and can also generate small alterations (Obe et al. 2002).

Chromosomal structural rearrangements can be classified as imbalanced or balanced, depending on whether there were loss or gain of DNA. Imbalanced rearrangements alter the quantity of genetic material in the affected chromosomes and thus is possible to disrupt normal gene balance, for example, deletions and duplications. In contrast, balanced rearrangements are copy number neutral events that change the chromosomal gene order and relative locations without overall net gain or loss of DNA, usually referring to reciprocal translocations and inversions (Griffiths et al. 1999). Note that these classifications are simply broad generalizations, since even seemingly balanced rearrangements could sometimes generate small deletions or duplications at the breakpoint regions as repercussion of the repair process (Young 2005).

1.2.1.1. Copy Number Variation

The term copy number variation (CNV) denotes a copy number change of DNA fragments sized 1 kb or above, be them duplicated and deleted. CNV often occur in regions reported to contain, or be flanked by, large homologous repeats (Freeman et al. 2006), and are an abundant form of variation in the human genome (Feuk 2010).

Duplication is a repetition of a genomic region, resulting in a gain of extra genetic material. A dupication usually occurs by unequal crossing-over between homologous chromosomes or sister chromatids, as well as abnormal meiotic segregation in a translocation or meiotic crossing-over in an inversion carrier (Czepulkowski 2001; Luthardt and Keitges 2001). A duplication is tandem if the duplicated segment is adjacent to the original copy, nontandem if it is nonadjacent, whether residing

in the same chromosome or inserted into another one, or terminal duplication if it affected the telomere (Bhat and Wani 2017).

On the contrary, a deletion is the loss of a DNA segment. Terminal deletions result from a single break within one chromosome arm with loss of material distal to the break, while interstitial deletions involve two breaks with loss of the material in between (Griffiths et al. 1999; Luthardt and Keitges 2001).

In general, duplications are less harmful than deletions due to the fact that a higher copy number of genes and regulatory elements are better tolerated by the cell then outright complete removal. Some duplications were even reported as potentially benign, but they may nonetheless be associated with some degree of phenotypical anomaly (Czepulkowski 2001). Additionally, larger CNV are often located in repeat-rich regions such as telomeres, centromeres and heterochromatin, away from high density protein-coding genes regions (Freeman et al. 2006). The degree of clinical severity is generally correlated with size of the duplicated or deleted segment (Luthardt and Keitges 2001; Griffiths et al. 1999).

Chromosomal structural variations, both balanced and imbalanced, may overlap with or locate nearby segmental duplications (SD), significantly more often than expected by chance alone. SD are blocks of DNA, ranging from 1 to 400 kb in length, present more than once in the genome and sharing over 90% sequence identity, found either on the same chromosome or on different, nonhomologous chromosomes (Freeman et al. 2006). Studies observed a significant association between the location of SD and regions of genomic instability that may lead to chromosomal rearrangements. The presence of large, highly homologous flanking repeats seems to predispose these regions to recurrent alterations (Sharp et al. 2005). As such, SD are sometimes found to be overrepresented near structural variation sites (Antonacci et al. 2009).

1.2.1.2. Balanced Rearrangements

Chromosomal translocations and inversions are generally balanced rearrangements that involve a break at two or more sites, and afterwards an incorrect reunion of the resulting fragments, without loss of genetic material. They occur with a frequency of 1 in every 2000 live births (Utami et al. 2014; Warburton 1991).

Translocations occurs when two nonhomologous chromosomes are each broken once, creating acentric fragments that later connect, forming a structure known as pachytene quadrivalent. When this temporary construction breaks off, their arms swap, trading places. Hence translocations are reciprocal in nature (Griffiths et al. 1999; Nambiar and Raghavan 2011). Inversions happen when two breakpoints occur in one single chromosome, and the region in-between reverses and reinserts into the opposite site (Griffiths et al. 1999; Young 2005). Chromosomal inversions will be further discussed in the next section.

1.3. On Chromosomal Inversions

1.3.1. Molecular Anatomy of an Inversion

Contrary to early predictions from classical cytogenetics, since their first identification in the 1920s inversions have been increasingly recognized as a relatively common source of variation (Alves et al. 2012; Kirkpatrick 2010). Inversions are indeed among the most common human constitutional karyotype anomalies detected in cytogenetic laboratories (Feuk 2010), found in about 2% of all humans and comprise approximately 10% of all structural rearrangements (Bhat and Wani 2017; Fickelscher et al. 2007; Muss and Schwanitz 2007). The frequency of *de novo* inversions identified is around 1 in 10,000 cases, with the risk of congenital anomalies in such cases being as high as 9.4% (Warburton 1991). Currently, as of February 2017, for humans, the National Center for Biotechnology Information's (NCBI) database of genomic structural variation (dbVar, https://www.ncbi.nlm.nih.gov/dbvar) reports 514 inversions validated by more than one experimental methods, such as FISH or microarrays.

Chromosome inversions, like other structural rearrangements, are often generated by non-allelic homologous recombination between inverted repeats or errors in the DNA repair mechanisms. The events create breaks that detach a segment of DNA and ultimately allow it to rotate itself before the repair mechanisms could enter once again in action to reunite the extremities (Puig et al. 2015b), as illustrated in Figure 1.1.

Large inversions are primary formed by non-allelic homologous recombination during the reparation process, where it involves regions of SD, since duplicated sequences can be inserted in an inverted orientation with respect to each other. However, for smaller inversions of under 10 kb the process remains uncertain, and more nucleotide-level information on breakpoints are needed to better understand the mechanisms and sequence motifs giving rise to small inversions (Feuk 2010).

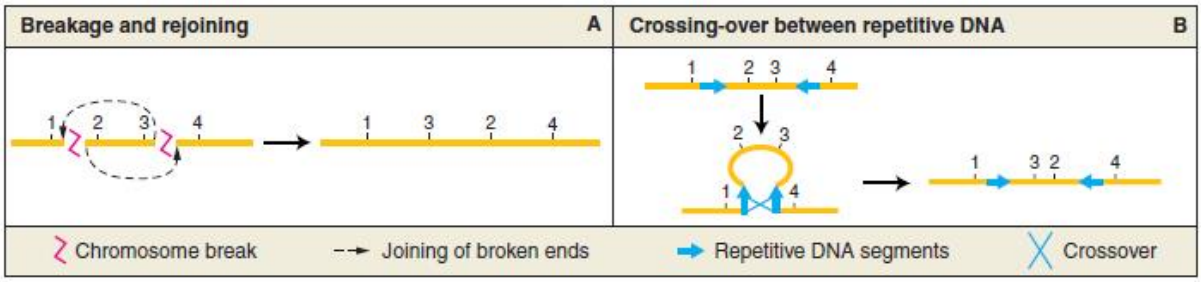

Figure 1.1. Basic mechanisms for the origin of inversions (adapted from Griffiths et al. 1999).
A) Chromosome break then rejoining
B) Crossing-over between highly repetitive regions
Chromosome regions are numbered 1 through 4.

Following an inversion event, the overall genetic behavior of the chromosome is determined by the location of the centromere relative to the inverted segment. If the centromere is inside the inverted segment then the inversion is said to be pericentric, whereas if it is outside then the inversion is paracentric (Griffiths et al. 1999).

Pericentric inversions can be detected through new arm ratios (Griffiths et al. 1999), but not paracentric inversions, which can only be detected microscopically by banding or by other chromosome landmarks if the inversion includes more than one band (Griffiths et al. 2000). Some breakpoints occur preferentially in certain regions, or hotspots, easing the detection of recurrent or suspected inversions (Muss and Schwanitz 2007).

During meiosis, a chromosome that houses an inversion follows one of two outcomes: either there is crossing-over between chromatids or there is not. If there isn't, then each resulting gamete will inherit a balanced rearrangement (Young 2005). But, if a cross-over does occurs between two nonsister chromatids within the inversion loop, like in the case of paracentric inversions, it will produce homologous centromeres in a dicentric bridge and an acentric fragment (Griffiths et al. 1999; Young 2005). As the chromosomes separate in anaphase I, the centromeres remain linked by the bridge, but tension will eventually break it at a random location, forming two chromosomes with terminal deletion. Since the acentric fragment cannot pair with its homologous due to its lack of a centromere, and the remaining fragment has two, both are unstable during mitoses and will be lost (Griffiths et al. 2000; Young 2005). This unviability means that paracentric inversions do not warrant indication for prenatal diagnosis, as the gametes containing deleted chromosomes are usually inviable because the crossing-over produces lethal products (Czepulkowski 2001), and are as such incompatible with life. In the contrary, pericentric inversion has no such problem with its single centromere. Still, they can produce recombinants with duplication and deletions, with its viability dependent upon the size of the unbalanced segments (Luthardt and Keitges 2001).

The decreased overall lethality and the comparatively ease of identification are the main reasons behind the higher detection frequency of pericentric inversions over paracentric inversions, 66% compared to just 34% (Luthardt and Keitges 2001).

Characterization of inversions represents a particularly remarkable challenge, especially in complex genomes, due to their balanced nature and breakpoints that are often located within highly identical inverted repeated sequences (Puig et al. 2015). Furthermore, cryptic imbalances can occur in half of individuals with a non-normal phenotype that harbors an alteration, increasing the overall complexity (Ordulu et al. 2016). A study conducted on 40 subjects with apparently balanced alterations, some of which are inversions, revealed cryptic rearrangements in nearly 40% of these cases (Higgins et al. 2008; Ordulu et al. 2016).

Due to their technical challenges, the study of inversions frequently fall out of favor for other easier to detect or clinically more severe or frequent variants, such as CNV and translocations. It means that much remain unknown about chromosomal inversions (Puig et al. 2015).

1.3.2. Polymorphic chromosome inversions

A structural chromosome anomaly is considered polymorphic if its allele frequency is above 1% in a given population (Freeman et al. 2006). In 1926, evidence of polymorphic inversions was found in *Drosophila melanogaster* (Corbett-Detig et al. 2012) and since then they were found to segregate in multiple species and are considered some of the most fascinating paradigms in evolutionary biology (Puig et al. 2015). Over time, knowledge about the genetic implication of polymorphic rearrangements in the human genome has grown rapidly, including that of inversions. The advent of

genome-scanning technologies has enabled the discovery of thousands of polymorphisms in phenotypically normal individuals (Antonacci et al. 2009).

Inversion polymorphisms are a universal phenomenon (Corbett-Detig et al. 2012). However, despite being present in virtually all species and extensively studied in model organisms, they are yet to be well characterized in humans (Cáceres et al. 2015) and a comprehensive map of inversions remains to be determined.

Because of complexity of the genomic regions where inversions typically occur (Cáceres et al. 2015) and the challenges related to the absence of affordable high-throughput methods, relatively few polymorphisms have been detected and characterized in humans (Antonacci et al. 2009), the majority falling within the 10 to 100 kb size interval (Alves et al. 2012).

Most of the currently known examples of polymorphic inversions came indirectly from studies of human diseases, where they have been identified from their association with susceptibilities to recurrent genomic rearrangements (Antonacci et al. 2009), or by their relationships with complex genetic disorders (Cáceres et al. 2012). Genome-wide single nucleotide polymorphism (SNP) data suggest that there could be hundreds of such inversions yet to be found in humans alone (Cáceres et al. 2015). As of February 2017, the invFest database (http://invfestdb.uab.cat/), which aims to store published polymorphic inversions in the human genome, reports 91 polymorphic inversion validated experimentally by direct observation, the majority of which are intergenic (Martínez-Fundichely et al. 2014).

To date, very few polymorphic inversions not directly related to pathology have been studied (Entesarian et al. 2008). Although not pathological, some of them seem to be connected to gene expression or even associated with certain phenotypes, like infertility and neurodegenerative diseases. Their mechanisms, however, are still not well elucidated (Puig et al. 2015a). One of the exceptions to the general obscurity is the well-studied polymorphic inversion located at 8p23, spanning 4.7 Mb and one of the largest know in humans (Salm et al. 2012), found in 26% of Europeans and 27% of Japanese individuals. Another example is the chromosome 4 inversion, of about 6 Mb in size and found in 12.5% of healthy controls (Feuk 2010). For most analyzed polymorphisms, there is no significant differences in the frequency of inversion alleles between ethnic groups (Antonacci et al. 2009), suggesting that these events have been occurring since ancient human history (Salm et al. 2012), showing a weak negative effect on reproductive fitness (Feuk 2010).

1.3.3. Inversions in chromosome 2

Chromosome 2 displays the highest recombination frequency for euchromatic pericentric inversions, at 11% (Muss and Schwanitz 2007). One of the most frequently observed inversion in humans is the familial inv(2)(p11q13), considered to be of no clinical significance (Feuk 2010) and is found in 0.1% of North Europeans (Entesarian et al. 2008). Over half of reported chromosome 2 inversions have their breakpoints located in p11.2 and q13 (Yakut et al. 2015)

It is estimated that one in every hundred newborns carries a pericentric inversion (Czepulkowski 2001). The incidence of pericentric inversions in chromosome 2 in the general population was reported to be between 0.0001% and 0.013%. Despite their large size, these inversions are usually

nonpathological. But, depending on its location, studies generally report some association with mental retardation, congenital malformations or reproductive failure (Djalali et al. 1986).

Inversions with clinical significance are not as frequently described in depth as other chromosomal rearrangements. This could be to the increased difficulty in characterizing inversions, especially those with cryptic rearrangements (Puig et al. 2015), as mentioned above. The pericentric inversion inv(2)(p23.3q24.3) was recently reported in an elderly individual with chronic thrombocytopenia and anemia (Kjeldsen 2015). And the pericentric inversion inv(2)(p15q21), described by Cohen et al. (1975), was considered of uncertain clinical significance in the observed Turner syndrome.

As of February 2017, in dbVar (http://www.ncbi.nlm.nih.gov/dbvar), there are only 24 validated chromosome 2 inversions reported in humans.

1.4. Phenotypical consequences of chromosomal anomalies

1.4.1. Congenital malformations associated with chromosomal rearrangements

Congenital malformations are defects in morphogenesis during embryogenesis that are identifiable at pregnancy or at birth. Although these malformations are often caused by genetic factors, environmental factors can also interfere with otherwise normal development (Corsello and Giuffrè 2012; Raymond and Tarpey 2006).

Chromosome anomalies are a major contributor of genetic diseases, resulting in heavy burden for public health systems worldwide (Corsello and Giuffrè 2012). They can cause an extensive array of clinical phenotypes such as infertility, congenital malformations syndrome, developmental delay, mental retardation and even autism spectrum disorder (ASD) (Raymond and Tarpey 2006; Vergult et al. 2014). It is very important to understand the molecular basis of these anomalies in order to improve diagnosis and treatment.

It is reported that the risk of congenital anomalies in newborns with apparently balanced chromosomal rearrangements is up to three times higher than in the general population, for which the risk of anomalies is between 2 and 3% (Warburton 1991).

These anomalies have been associated with over 60 identifiable syndromes, and are detected in 50% of spontaneous abortions, 6% of stillbirths, 5% of couples with two or more miscarriages and approximately 0.5% of newborns. The frequency of anomaly increases with the mother's age, and for women over 35 years old it reaches 2% of all pregnancies (Le Caignec et al. 2005; Luthardt and Keitges 2001). Furthermore, the risk for congenital malformations is even higher in consanguineous unions (Leonard 2016).

A starting event, like the occurrence of alteration or mere random accidents during the formation of reproductive cells or in early fetal development, may originate a series of dysmorphogenetic processes producing a cascade of defects and malformations. The severity of these malformations depends on the nature, size and location of original event and how it influenced the downstream processes (Corsello and Giuffrè 2012). These changes in chromosome structure due to rearrangements can cause chromosomal disorders that can be inherited (GHR 2014). Studies observed that familial transmission

of chromosomal rearrangements is mainly through female carriers (Batista et al. 1994), though the reason behind this phenomenon is unclear.

Studies report that chromosomal structural anomalies can produce similar phenotype to those caused by point mutations, if directly disrupted important genes or by some position effect as consequence of rearrangement, in some cases easing their analysis (MacIntyre et al. 2003). Whenever a balanced rearrangement occurs in a non-coding region, however, predicting pathogenic consequences can become very challenging if the evaluation focuses only on the nearest genes (Ordulu et al. 2016). Also, since balanced rearrangements change the genomic localizations and order of genes, they can have profound effects on their expression, even if they remained intact (Avelar et al. 2013). They can also modify gene pattern activity by dissociating them from its original regulatory elements and/or placing it under the control of other regulatory elements (David et al. 2013; Mihelec et al. 2008; Puig et al. 2015). Because of all the positional changes, it is recommended to analyze not only disrupted gene-regulatory elements pairs but also interaction that may be happening between the primary alteration and any other alterations.

Phenotypically normal carriers of balanced rearrangements can produce an offspring with imbalance, especially if it involves a large region (Czepulkowski 2001). As a general rule, any degree of chromosome imbalance involving one or more of the autosomes will have serious adverse effects (Young 2005). Still, even an apparently balanced alteration can be associated with severe malformations. About 6% of *de novo* balanced rearrangements detected at amniocentesis are associated with an abnormal phenotype (Tabet et al. 2015), and many were linked to clinical features like psychomotor developmental delay, dismorphism, defects of heart formation and ASD (Raymond and Tarpey 2006; Utami et al. 2014).

In individuals with ASD, chromosomal anomalies were detected in 7.4% of cases (El-Baz et al. 2016), some of which known to be implicated in the phenotype (Tabet et al. 2015). Actually, the large number of genes identified as associated with complex clinical features, like psychomotor developmental delay, suggests that it should be the common consequence of many affected cellular processes and that no single mechanism is likely to be the only cause of phenotype (Raymond and Tarpey 2006). Considering that the rate of chromosomal rearrangement in people with mild learning disability may be as high as 19% (MacIntyre et al. 2003), if the alteration is in some way linked to disability, the number of possible molecular causes are certainly very high.

1.4.2. Clinical Implication of Inversions

Biologically, even large inversions are likely to be neutral, without obvious phenotypic consequences (Feuk 2010). Studies have also shown that both carriers and non-carriers of inversions have a similar rate of miscarriages and death in the neonatal period (Muss and Schwanitz 2007), despite the selective forces against excessive harm due to genomic changes early in development (Kirkpatrick et al. 2011).

Families carrying only small and sub-clinical inversions may transmit the aberrant chromosome through several generations and never be detected, as there may be no reason for cytogenetic screening. Among healthy offspring of inversion carriers, there is an expected 1:1 ratio of them being inversion carriers to noncarriers (Honeywell et al. 2012). Nonetheless, the majority of carriers of

inversions have no clinically significant phenotype, both in the heterozygous and in the rare homozygous state (Muss and Schwanitz 2007).

Carriers of balanced rearrangements are usually phenotypically normal if no essential genes are affected by the breakpoints or if it does not fall between a gene and its transcription regulatory elements. (Feuk 2010; Luthardt and Keitges 2001). Since inversions typically do not result in changes in copy number, only of orientation, they could appear to be neutral variants without phenotypic effects of clinical significance. However, in reality, some inversion are far from harmless. For example, they are potential mechanical causes of subfertility (Alves et al. 2012) and another well-known example is the disease-causing recurrent inversion affection 3' region of the *F8* (exon 22 to 26) is found in 43% of patients with severe Hemophilia A (Feuk 2010).

If one of the breakpoints is within or closely adjacent to a gene of essential function, then it could act as a lethal gene mutation and result in a pathology (Griffiths et al. 2000; Muss and Schwanitz 2007), but it still highly depends on the gene's role and the mechanisms acting on it. Genes within an inversion can be entirely unaffected, unlike those within CNV, which are always affected by a dosage imbalance (Feuk 2010).

However, compared to the downstream effects of CNV, the genes disrupted by inversions are presumably more severely affected, due to increased risk of protein truncation or heterozygous inactivation of the affected allele (Utami et al. 2014). For example, an inversion truncates the X-linked *IDS* gene in 13% of Hunter syndrome patients, potentially causing or exacerbating the disease (Puig et al. 2015b). Very rarely, it is possible that the rejoining of two partial, truncated genes can originate new protein or promoter, in other words, gene fusion. Depending on the resulting products, they might be pathogenic, neutral or even benign, giving rise to new functions (Griffiths et al. 1999; Kloosterman and Hochstenbach 2014).

Heterozygous inversion generally are not linked to disease nor do the carriers show adverse phenotype. But the carriers produce the expected array of abnormal meiotic products, and often also have mechanical pairing problems in the region of the inversion, reducing the frequency of crossing-over and leading to an increased risk for chromosomally abnormal offspring (Griffiths et al. 2000; Luthardt and Keitges 2001). An inversion cannot be bred to homozygosity if it is lethal, but it may still be viable if only nonessential genes were affected (Griffiths et al. 1999). In such cases, the inverted chromosomes pair and crossover normally, the meiotic products are viable, and the linkage map will show the inverted gene order (Griffiths et al. 2000). While inversions with an apparent negligible phenotype may appear of little significance, by carrying a certain allele, other rearrangements in the same genomic region can be predisposed, which in turn could lead to disease. In these cases, the parents of affected individuals would show a higher frequency of one of the inversion alleles compared with the general population (Puig et al. 2015b). For example, direct association between an inversion carrier status, including common polymorphic variants, and increased risk for deletions in the offspring has been established in several microdeletion syndromes (Feuk 2010). Nonetheless, the probability for an inversion carrier to predispose a child to other rearrangements is still extremely low (Puig et al. 2015b).

Aside from the recognized problems that hinders rearrangements detection and the analysis of their impact, the knowledge on the contribution of structural variants to disease is possibly being underestimated owing to the use of exome sequencing to identify causal genes: this sensitive method fails to detect genes truncated by breakpoints (Puig et al. 2015b).

1.5. Identification of Chromosomal Anomalies

1.5.1. Conventional Technologies

Classical study of chromosomes has been based primarily on large-scale rearrangements via karyotype analysis with classical G-banding techniques (Alves et al. 2012). Only asymmetrical forms which give rise to acentric fragments or extensive alterations across several chromosomal bands are readily visible (Obe et al. 2002), and even significantly larger rearrangements may escape detection if the affected segments lead to little difference in the banding pattern. Chromosomal aberrations detected through these methods are of very low resolution, in the magnitude of 5 to 10 Mb, aside from being laborious and not permitting a global unbiased discovery (Feuk 2010).

As consequence, subsequent fluorescence *in-situ* hybridization (FISH) mapping across the breakpoints was used for a more detailed delineation of these events (Utami et al. 2014), with resolution between 200kb and 2MB depending on the type of FISH used. Also, with the aid of differential fluorochrome mixes, it is possible to digitally "paint" the chromosomes (pseudocolor) and thus follow the exchange of segments between them. However, it generally is not applied to inversions nor other alterations within the same chromosome as the kits usually "paint" each chromosome with one color (Obe et al. 2002).

Similarly, bacterial artificial chromosomes spanning the chromosome breakpoint can be used for segregation pattern analysis (Bhatt et al. 2007). For inversions larger than 2 Mb, inversion breakpoints can be mapped by metaphase FISH using two probes located more than 2 Mb apart inside the inverted region. Smaller inversions of down to 200 kb can be resolved by three-color interphase FISH. Still, it is difficult to correctly assess the probes relative position, especially when breakpoints map to duplicated sequences (Antonacci et al. 2009), aside from remaining target-based (Alves et al. 2012).

More recently, array comparative genome hybridization (array CGH) was developed. These genomic microarrays with high density of SNP markers are extremely capable in the identification of copy number changes or imbalances genome-wide at a relatively high resolution, being conventionally used for the detection of duplications or deletions as well as loss of heterozygosity (Feuk 2010). Actually, nowadays, it is one of the preferred clinical diagnostic tests to detect anomalies in prenatal diagnosis, when one or more major structural abnormalities were identified in a fetus by ultrasonograph (Ordulu et al. 2016) or for individuals with multiple congenital anomalies and developmental delays (Leonard 2016). However, this method fail to identify copy number neutral rearrangements like inversions because they are unable to determine genomic positions, only of copy number state (Utami et al. 2014), and are prone to artifacts like calling of false positives or high background noise (Alves et al. 2012; Curtis et al. 2009).

The characterization of structural variants are indeed often stalled by technical challenges in genome-wide screening (Pinto et al. 2007). Detection and analysis of rearrangements with low resolution technologies can be difficult or inconclusive (Obe et al. 2002). Additionally, evidence show that repeated sequences are common at rearrangements breakpoints, making characterization by standard molecular approaches even more challenging (Antonacci et al. 2009).

Since conventional cytogenetic often overlooks complex anomalies, higher resolution platforms must be called into action (Gregori et al. 2007).

1.5.2. The advent of Next-Gen Technologies

1.5.2.1. Sanger Sequencing

The era of "first-generation sequencing" started with Sanger sequencing. Since its conception in 1975, by Edward Sanger, it has been used extensively and is nowadays still regarded as the gold standard for the determination of nucleotide sequences (Grada and Weinbrecht 2013). Among its many notable achievements, it permitted the demystification of previously believed to be near-impossible feats, such as the sequencing of an entire complex genome, the Human Genome Project, completed in 2003 (Metzker 2010).

There are two technical approaches in Sanger sequencing, depending on whether it is a *de novo* sequencing or a resequencing. In the first case, DNA is randomly fragmented, cloned into a plasmid then used to transform *Escherichia coli*; in the second, a PCR amplification is performed. The sequencing takes place during cycles of template denaturation, primer annealing and extension. During extension, a fluorescently labeled dideoxynucleotides (dNTP) is incorporated then stochastically terminated. Inside a capillary based polymer gel, laser excitation of end-labels permits the identification of the nucleotide identity based on its signal. Software translates the signal and generates chromatogram. Sanger sequencing can reach read-length up to of around 1000 bp with over 99.9% accuracy (Shendure and Ji 2008).

1.5.2.2. Next-Generation Sequencing

The demand for the development of cheaper and faster sequencing to generate large amounts of data has increased greatly over the years, driving the development of new technologies. Subsequently, in recent years, there was a slow yet steady shift away from Sanger sequencing - the time of the "second-generation", also known as Next-Generation Sequencing (NGS), has arrived (Metzker 2010).

The major advance offered by NGS is the ability to produce an enormous volume of data in a relatively short amount of time (Metzker 2010). While NGS is still indeed expensive, novel approaches and refinements of existing methods are reducing the cost per base while increasing the throughput, easing the elucidation of variations not covered by other methods. Furthermore, the cost per base is significantly lower when compared to Sanger sequencing, due the sheer amount of bases identified at once. This capacity of sequencing many millions or billions of bases in hours is something that Sanger sequencing cannot even come close to match at its current form (Pettersson et al. 2009).

Many studies have already ventured into this new promising land. Those that applied NGS-based methodologies into their work commend for its usefulness and effective ability to provide a vast amount of information, with an incredibly fine detail and a consistent validation rate, enabling the identification of the causal gene for many rare Mendelian disorders (Hanscom and Talkowski 2014; Talkowski et al. 2011; Utami et al. 2014; Vergult et al. 2014).

Different NGS platforms coexist in the marketplace, each having specific advantages and disadvantages for particular applications (Metzker 2009). This has made sequencing accessible to

more laboratories, rapidly increasing the amount of research and clinical diagnostics being performed (Grada and Weinbrecht 2013).

Every NGS platform embodies a complex interplay of chemistry, hardware and software engineering, allowing a highly streamlined sample preparation steps prior to DNA sequencing (Mardis 2008). Having been developed by different manufacturers, each platform use propriety methods that slightly differs from one other, with Illumina's currently being the most widely used (Metzker 2010). Nonetheless, they share similarity in how the nuclear workflow is performed.

The overall NGS methodology include template preparation, sequencing and imaging and, finally, data analysis (Grada and Weinbrecht 2013). Template preparation consists of building a library of nucleic acids. This is accomplished by random fragmentation of DNA, followed by a ligation of adaptor sequences and finally amplification of the library in preparation for sequencing (Mardis 2008; Shendure and Ji 2008). The sequencing process itself consists of alternating cycles of enzyme-driven biochemistry and imaging-based data acquisition. Most widely used platforms rely on sequencing by synthesis, in which the library fragments acts as template. The sequencing occurs through a cycle of washing and sequential flooding of the fragments with known nucleotides. As nucleotides incorporate into the growing DNA strand, they are detected and recorded sequentially (Grada and Weinbrecht 2013). The captured signal is then analyzed with bioinformatics tools, where the obtained sequences are aligned to a known reference sequence (Metzker 2010). In sum, these technologies have the advantages of *in vitro* construction of a sequencing library, followed by *in vitro* amplification to generate sequencing features (Shendure and Ji 2008).

The most comprehensive application of NGS may be the resequencing of human genomes to enhance our understanding of how genetic differences affect health and disease (Metzker 2010). Various studies have described the use of whole-genome sequencing to identify genomic breakpoints of chromosomal rearrangements at nucleotide resolution, facilitating candidate genes identification (Kloosterman and Hochstenbach 2014; Utami et al. 2014; Vergult et al. 2014).

1.5.2.3. Mate-pair sequencing

To define rearrangement breakpoints, studies have increasingly resorted to paired-end sequencing libraries. Introduced in 2007, it has since been used to identify more than a thousand structural variations in the human genome at unprecedented resolution. It involves the mapping of pairs of sequence reads to a reference genome, derived from the two ends of a single DNA segment (Kloosterman and Hochstenbach 2014).

A major breakthrough in the discovery of structural rearrangements came in the form of mate-pair sequencing, a type of paired-end sequencing which resorts to larger segments of 2 to 6 kb between each read pair (Talkowski et al. 2011; Utami et al. 2014; Vergult et al. 2014). With reads so far apart, mate-pair sequencing is able to cover difficult DNA sequence features such as repetitive regions. These libraries are particularly useful for the analysis of inversions, since the presence of inverted reads at breakpoints hinder their identification using single reads or paired-end sequencing of shorter fragment size (Aguado et al. 2014). Many of the predicted inversions so far have used this strategy (Puig et al. 2015b). Furthermore, Hillmer and colleagues (2011) have demonstrated that large insert fragment sizes provided higher physical coverage with minimum sequencing efforts.

Large-insert library preparation proposed by Talkowski and collaborators (2011) is a type of library prepared for whole-genome mate-pair sequencing. It is a customized library preparation optimized for structural rearrangements. It is also generally less expensive than commercially available counterparts. The main construction steps of this mate-pair library are schematized in Figure 1.2., which describes the process of creating sequencing-ready libraries for Illumina platforms.

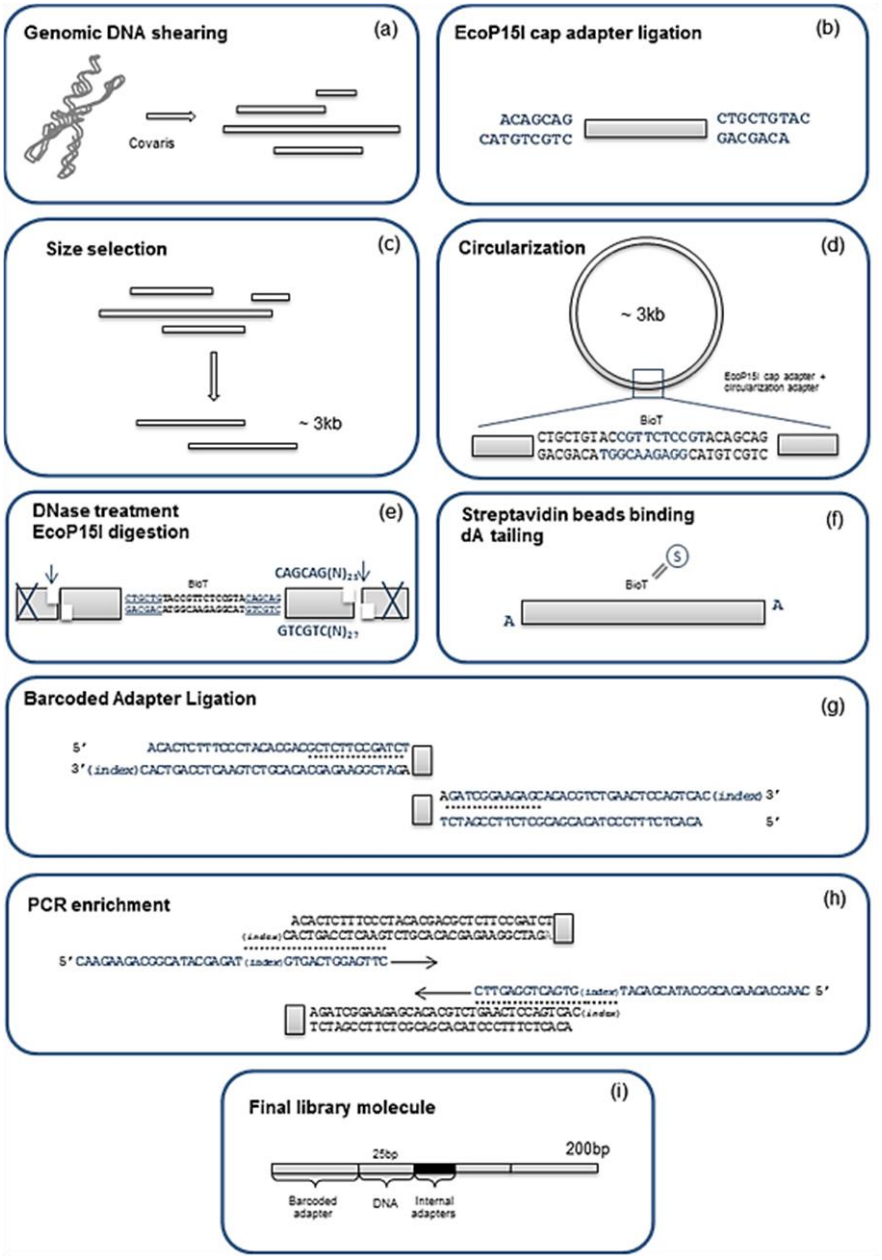

Figure 1.2 Large-insert library preparation for NGS, as described by Talkowski et al (2011).
A) Genomic DNA sheared using an ultrasonicator to a mean size of 3 kb.
B) Cap adaptors containing a EcoP15I recognition site are ligated
C) DNA size-selection of fragments of between 2.5 and 6 kb
D) Circularization of DNA with biotinylated adapters
E) DNase treatment to remove linear DNA, then digestion with EcoP15I restriction enzyme on the circularized DNA
F) DNA fragments bind to streptavidin beads, and dA tailing
G) Ligation of barcoded adapters with Illumina specific indexes enables multiplex during sequencing; dots indicates where the oligos anneal
H) PCR enrichment of the library, dots indicating annealing of primers and arrows the direction of extension during PCR
I) Final library molecule of around 200 bp, ready for sequencing.

1.5.2.4. Analysis of sequencing data

After obtaining the final library fragments, the ends are sequenced in a massive and highly parallel manner. The resulting reads then mapped to a reference genome. The majority of the read pairs would map concordantly to a reference genome, as expected. However, if the read pairs do not map concordantly, in other words, if a cluster of several read pair maps to highly distant sites or with an unexpected orientation, then it points to the existence of a structural rearrangement in between. Abnormal location and orientation mappings of the read pairs are indicative of the presence of an inversion breakpoint (Feuk 2010; Puig et al. 2015b; Utami et al. 2014), as illustrated in Figure 1.3.. Furthermore, identification of split read mapping gives insight into the precise breakpoint junction sequence (Kloosterman and Hochstenbach 2014).

This technology is capable in identifying rearrangement breakpoints at nucleotide level. Generally, however, while the depth of coverage is in average 60x, it varies highly from region to region. In locations where the average is low, it may not be enough for the identification of structural chromosomal anomalies with a nucleotide resolution. Thus, it usually predicts breakpoint coordinates down enough to enable Sanger sequencing (Utami et al. 2014).

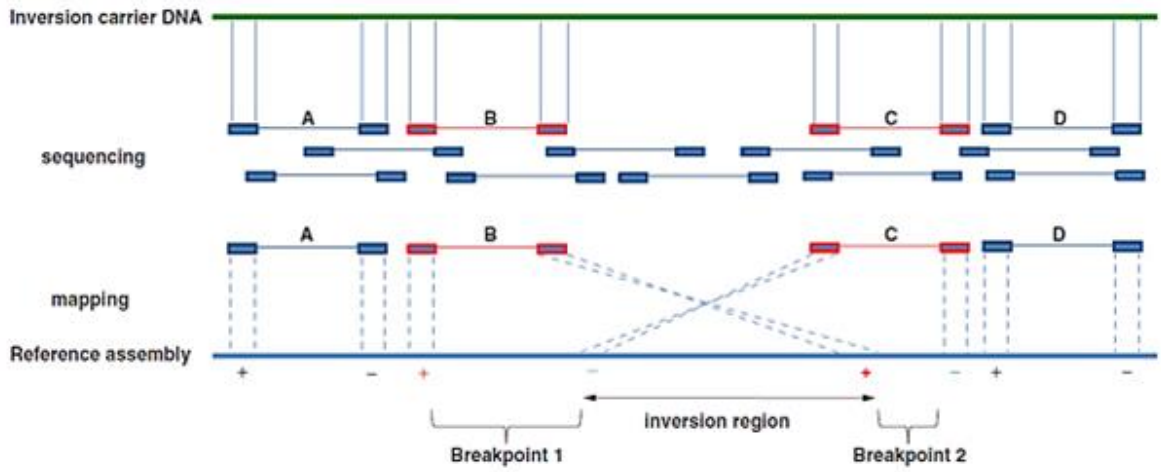

Figure 1.3. Inversion breakpoints discovery using mate-pair sequencing and mapping (adapted from Feuk 2010).
Lines in red show fragments spanning into the inversion while blue are the normal unaltered fragments, with the boxes being read-pairs. Orientation of reads that map concordantly to reference genome is indicated by "+", otherwise by "-". When the pairs align to the reference genome, the majority map in a +/- orientation (A and D), since they are in a region without any structural alteration. In the region containing an inversion, the breakpoints would map as +/+ and -/-, respectively (B and C).

In the end, while this approach is appealing because it is validated in fine-mapping structural breakpoint and for interrogation of structural polymorphisms, it still cannot be used as the sole methodology. For example, it is very challenging to distinguish an identified inversion breakpoints from duplications that reinserted in an inverted orientation (Corbett-Detig et al. 2012). Additionally, as mate-pair sequencing heavily relies on the reference assembly, very rare or unique alleles and any mis-assembly are serious challenges to overcome. The small size of the library fragments may be problematic for alignment to regions with high identity SD or with repetitive sequences (Feuk 2010; Rasekh et al. 2015).

In other words, although sequencing technologies have certainly proven their worth as the most efficient methods in chromosomal rearrangements screening and breakpoint identification, overlapping various methodologies remains the best practice worldwide (Ordulu et al. 2016). With a boarder approach, the various existing methodologies would be able to cover for each other's weak points.

1.6. Objectives

The principal objective of the present study is the identification of molecular alterations and of candidate genes responsible for a malformation syndrome in an individual with an apparently balanced maternally inherited pericentric chromosome inversion inv(2)(p16.1;q14.3)mat. More specifically, the goals involve:

- Identification of inversion breakpoints with nucleotide resolution
- Characterization of the inversion breakpoint regions
- Clarification on the role played by the detected inversion in the malformation syndrome
- Investigate existence of other chromosomal structural anomalies
- Conduct familial segregation analysis on additional family members
- Expression profiling of cell lines established from the proband

2. MATERIALS AND METHODS

2.1. Sample Preparation

2.1.1. Biological samples

Peripheral whole blood from the proband and his mother were collected in 2007 after informed consent. The blood was stored in specialized collection tubes containing ethylenediamine tetraacetic acid (EDTA) for downstream genomic DNA extraction, and in collection tubes with heparin sodium for establishment of cell lines.

In 2017, additional biological samples were obtained. Peripheral whole blood from the proband, his mother and his father were collected.

2.1.2. Cell culture

Lymphoblastoid cell lines (LCL) are commonly used as source of difficult to obtain biological material. The somatic mutation rate of LCLs is estimated to be 0.3%. Additionally, the relative ease of cell maintenance makes them an attractive alternative of genetic material source (Sie et al. 2009).

Cell culture procedures were done in accordance to the described in the papers published by David et al. (2009; 2013).

The LCL from the proband was placed in culture for the continuation of this case study. The cells were maintained in complete cell medium developed for suspension cells. The cell medium was prepared as follow: Roswell Park Memorial Institute medium (Gibco), 15% heat inactivated fetal bovine serum (FBS-Hi) (Gibco), 1% glutamine and 1.5% of Penicillin-Streptomycin. The LCL was in a constant temperature of 37ºC, in an incubator with dioxide carbon levels at 5%. Every two days, the LCL was observed to confirm their health and growth. The cells were counted often to maintain optimal cell concentration. Trypan blue were used to selectively staining dead cells. A Neubauer chamber was used and cells counted under a microscope, where only live (non-stained) cells were considered. An ideal concentration would be around $0.7x10^6$ cells per ml, generally in 10 ml of cell medium, for a total cell count of $0.7x10^7$.

2.1.2.1. Cryogenic preservation of LCL

The LCL were cryogenically preserved in liquid nitrogen (-170ºC) as to retain a backup of biological sample. The viability cells greatly depends on how well the process underwent. If the procedure was not done with care, the cells may take a long time to recover or even die off (Withers and Street 1977). In brief, the cells were transferred to a cryovial in a cold freezing medium prepared with 90% FBS-Hi and 10% dimethyl sulfoxide. The tubes were kept 24 hours at -80ºC in a cryobox filled with isopropanol, allowing for a constant rate of temperature drop (1ºC/min). Afterward, they were stored in a container with liquid nitrogen.

2.1.2.2. Cells pellet preparation for DNA and RNA extraction

For the purpose of DNA and RNA extraction from LCL, dry cell pellets were prepared. Briefly, the entire culture volume was transferred into a 15 ml centrifuge tube, centrifuged and washed twice, then transferred to a 2ml microcentrifuge tube and centrifuged once again. The cell pellet was air dried, then stored at -20ºC.

2.1.3. DNA extraction for peripheral blood and LCL

Genomic DNA from peripheral blood samples was extracted using the QIAamp Midi DNA Blood kit (QIAGEN). It was designed for the extraction and purification of genomic DNA up to 50 kb from blood samples. The procedure was based on the manufacturer's protocol (QIAamp DNA Blood Midi/Maxi Handbook, http://www.qiagen.com/no/resources/resourcedetail?id=bf32146a-77fd-40c2-8743-c28974f7935b).

In brief, peripheral blood was incubated at room temperature for 2 hours. Proteinase K (PK) and Buffer AL was added to 1.2 ml of blood, then vigorous mixed and incubated at 65ºC for 30 minutes to promote cell lysis. The DNA were precipitated with molecular biology grade absolute ethanol. Wash buffers AW1 and AW2 removed traces of contaminants like proteins aided by a series of centrifugations. The DNA was then eluted in low TE buffer (10mM Tris pH7.5, 0.1mM EDTA), which minimizes the risk of downstream enzymatic reactions inhibition due to EDTA and DNA degradation of elution in pure water.

The Blood & Cell Culture DNA Midi kit (QIAGEN) was used for the extraction of genomic DNA from LCL, according to manufacturer's instructions (QIAGEN Genomic DNA Handbook, http:// www.qiagen.com/ie/resources/resourcedetail?id=402bb209-4104-4956-a005-6226ff0b67d5). This kit is suitable for the extraction of high molecular weight genomic DNA of up to 150 kb, with the downside of being a long and laborious process.

In summary, $2x10^7$ LCL cells, dry pelleted as described above, was resuspended. Buffer C1 and ice-cold water was used to lyse the cells, keeping the nucleus intact. Nucleus disruption and protein denaturation were promoted by Buffer G2 and PK with an incubation at 50ºC for 1 hour. The DNA then were bound to the column while other cell constituents gently passed through by gravity flow, followed by several wash steps to remove any remaining contaminants. The purified DNA was eluted in Buffer QF (1.25 M NaCl, 50 mM Tris·Cl pH 8.5, 15% isopropanol).

The extracted DNA was quantified with Nanodrop ND-1000 spectrophotometer (ThermoFisher), using the corresponding elution buffer as blank, according to manufacturer's instructions (NanoDrop 1000 Spectrophotometer V3.8 User's Manual, http://www.nanodrop.com/Library/nd-1000-v3.8-users-manual-8%205x11.pdf). The DNA quality was accessed with the absorbance ratios A260/A280 and A260/230, where a value of 1.8 and 2.0, respectively, indicates a sample free of contaminants such as phenol and proteins. Note that spectrophotometers detects a variety of nucleic acids, so if a DNA sample is contaminated with RNA, it would overestimate the overall quantity (Desjardins and Conklin 2010).

To confirm the quality control with Nanodrop, as well as to check the integrity of the DNA, agarose gel electrophoresis was used. Lambda DNA/*Hind*III marker was used as both quantity and molecular weight reference. As such, 100 ng and 200 ng of Lambda DNA/*HindIII* and DNA samples were loaded into a 0.8% agarose gel in TAE buffer and ethidium bromide, and ran for 3 hours at 45 V. A high quality DNA would appear as a high molecular weight and distinct band with little to no smear. Quantity-wise, it can be inferred by comparing the band intensity of the DNA and the marker's – the quantity would be similar if the intensity is nearly the same.

2.1.4. RNA extraction from LCL

The LCL was harvested as described above. Total RNA extracted with QIAamp RNA Blood Mini Kit (QIAGEN), in accordance to the manufacturer's indications (QIAamp RNA Blood Mini Handbook, http://www.qiagen.com/us/resources/resourcedetail?id=5ea61358-614f-4b25-b4a5-a6a715f9d3aa).

In brief, $0.8x10^7$ cells were resuspended and disrupted with Buffer RTL, beta-mercaptoethanol added to reduce RNase activity, and then the lysate transferred to a shedding column. The freed RNA was precipitated with molecular grade absolute ethanol and bound to a column membrane. Deoxyribonuclease I was added to remove any existing DNA molecule, and sequential washes removed remaining cell debris and other contaminants. The RNA was eluted in the provided nuclease-free water and quickly stored at -80ºC.

The extracted RNA was immediately quantified with a Nanodrop ND-1000 spectrophotometer (ThermoFisher), using nuclease-free water as blank, according to manufacturer's instructions. The absorbance ratios A260/A280 and A260/A230 must fall between 1.8 and 2.0 to be considered "pure" samples free of significant contaminants.

A significantly more sensitive, albeit slower and more expensive, method of quality control of RNA samples is the use of 2100 Bioanalyzer (Agilent). This control was done by an external service provider. Agilent's RNA 6000 Nano Kit was used, according to manufacturer's instructions. Due to the sensitive nature of RNA, only samples with a RNA Integrity Number of at least 8.0 were used for downstream applications.

2.2. Polymerase Chain Reaction

2.2.1. Primers Design

Sequencing results from NGS allowed for the determination of the chromosome 2 inversion breakpoints, delimitating the region within which a breakpoint is located. As such, primers were designed on each side of this region, with the software OLIGO Primer Analysis and NCBI's Primer-BLAST tool (http://www.ncbi.nlm.nih.gov/tools/primer-blast). The first calculates free energy, hybridization temperature and nucleotides secondary structure (Rychlik and Rhoads 1989). The second designs primers specific to a PCR target, aligning them against a reference genome to minimize non-specific amplifications (Ye et al. 2012).

In order to increase the probability of obtaining a high quality primer, candidate primers had to be confirmed to be in the correct position and have their specificity checked using BLAST against human reference genome. Furthermore, primers located in highly repetitive regions were avoided whenever possible. Of the possible primers for each side of the breakpoint region, the one that had the lowest score for dimer formation and self-complementarity, given by OLIGO software, was chosen (Table 2.1.). The primers were synthesized by an external provider.

Table 2.1. Primers for the chromosome 2 inversion.

Fragment	Designation	Primer sequence (5'-3')	[a] Primer localization	[b] Anneal T (ºC)	Primer size (bp)
inv(2)(p16.1) junction fragment	AC015982-1F	GGGCCAACTGGATAACTAAAAA	chr2:55,934,504-55,934,525	65.2	22
	AC096744-3F	AAAAAGAGCAAAGTTGGGAGCA	chr2:123,767,376-123,767,397	67.9	22
inv(2)(q14.3) junction fragment	AC015982-2R	TTTAAGGAGCAAGAGAACACGTT	chr2:55,935,344-55,935,366	65.1	23
	AC096744-4R	TCTAGTTCACATCCTTTGCCCA	chr2:123,768,000-123,768,021	66.8	22

[a] Numbers indicate the position of primers in the human genome assembly GRCh37.
[b] Annealing temperature given by OLIGO software.

2.2.2. Amplification of DNA by PCR

The inversion breakpoint junction fragments were amplified by polymerase chain reaction (PCR), using the designed primers, to determine the breakpoints' genomic position.

For chromosomal inversions, each pair of primers used for amplification of junction fragment is of the same orientation (i.e. primers forward with forward, reverse with reverse), as it is composed by a union of inverted and non-inverted sequences. The breakpoint at inv2p16.1 used primers AC015982-1F (as forward) and AC096744-3F (as reverse), while the breakpoint at inv2q14.3 used primers AC015982-2R (as forward) and AC096744-4R (as reverse), as described in Table 2.1..

For the PCR reaction 50 µl of mixture was used. The master mix was prepared as follow: per reaction, 50 µl of in-house premix (20 µl dNTP 100 mM; 1 ml GeneAmp 10X PCR Buffer I (Applied Biosystems); 9 ml water treated with diethylpyrocarbonate (Bioline), 0.5 µl of forward primer (300 ng/µl), 0.5 µl of reverse primer (300 ng/µl) and 0.5 µl of AmpliTaq DNA Polymerase (Applied Biosystems). Afterwards, 50 µl of master mix was transferred into 200 µl thin-walled microcentrifuge tubes then, finally, 100 ng of genomic DNA was added, except for the negative controls.

The amplification was performed in a thermalcycler (Biometra), with the following setup: an initial denaturation of the DNA at 94ºC for 3 minutes, followed by 35 cycles of a 45 seconds long denaturation step at 94ºC to create single-stranded templates, 40 seconds of primers annealing to the templates at 63ºC for the inv2p16.1 breakpoint and 61ºC for the inv2q14.3 breakpoint, and an extension at 72ºC for 75 seconds. Afterwards, a 3 minutes final extension guaranteed that the DNA molecules had enough time to be synthesized and all single-stranded DNA was fully extended. The reaction was then cooled down to 4ºC to avoid degradation to due extended exposure to heat.

2.2.3. Quality control of PCR products

PCR amplification success was confirmed by standard horizontal agarose gel electrophoresis. For that, 5 µl of PCR product was mixed with in-house orange-G loading buffer, then transferred to a 1.2% agarose gel in TAE buffer 1x. For molecular weight marker, Hyperladder 50 bp (Bioline) was used. The electrophoresis ran for 90 minutes at 50 V.

If the amplification proceeded with sufficient efficiency, then a distinct high intensity band within the expected size range would be seen. And if the primers were specific enough for the sequence of interest, no other bands aside from the high intensity band would appear on the gel.

2.2.4. Purification of PCR products

The PCR products amplified successfully were purified with Amicon Ultra-0.5 ml Centrifugal Filters (Millipore), following the manufacturer's instructions (Amicon Ultra-0.5 Centrifugal Filter Devices User Guide). This kit is designed for the removal of small nucleic acids like primers and unused dNTP, as well as enzymes and salts. Filters in the column retain higher molecular weight DNA while the contaminants flow through it by centrifugation. By diluting the PCR product with low TE buffer to the maximum volume of 500 µl, it increased the efficiency of contaminants removal. The sample recovery was done with centrifugal force by inverting the column into a clean microcentrifuge tube.

The purified sample was quantified with Nanodrop ND-1000 spectrophotometer (ThermoFisher), according to manufacturer's instructions.

2.3. Sanger Sequencing

Sanger sequencing The ABI PRISM BigDye Terminator Cycle Sequencing Kit (Applied Biosystems) was used to prepare the DNA samples for sequencing.

For that, the sequencing reaction was prepared by mixing 3 µl of Sequencing Buffer and 2 µl of Big Dye then 40 ng of the previously purified and quantified PCR product was added, followed by 1 µl of one of the primers (10 ng/µl) for that fragment, either forward or reverse, is used in one reaction – in other words, each PCR product was sequenced twice, each with a different primer used in the original PCR amplification. Finally, nuclease-free water was added to bring the total volume to 20 µl. The reaction was then conducted in a thermalcycler as following: a denaturation step at 96ºC for 10 seconds, an annealing step at 61ºC for the inv2p16.1 fragment and 59ºC for the inv2q14.3 fragment (2ºC lower than at PCR amplification) for 5 seconds and an extension step at 60ºC for 4 seconds. A final extension was performed also at 60ºC, for 3 minutes. The reaction was then put on hold at 4ºC.

The Sanger sequencing was performed by an external service provider, on an ABI PRISM automatic sequencer (Applied Biosystems). The results were returned in form of chromatograms, which were read against the reference genome. A breakpoint is found when the sequenced fragment stops matching the reference sequence. Also, any other small alterations, such as SNPs or microdeletions, are recorded and investigated on public databases.

2.4. Genomic Array

Genomic DNA from the proband and his parents were analyzed by high-resolution CytoScan HD array from Affymetrix. This array excels in the identification of CNV, such as duplications and deletions, as well and other changes like loss of heterozygosity, and 96% of known genes are represented (Uddin et al. 2014). The assay was performed according to the CytoScan HD assay user manual (http://tools.thermofisher.com/content/sfs/manuals/cytoscan_assay_user_manual.pdf). Figure 2.1. summarizes the process.

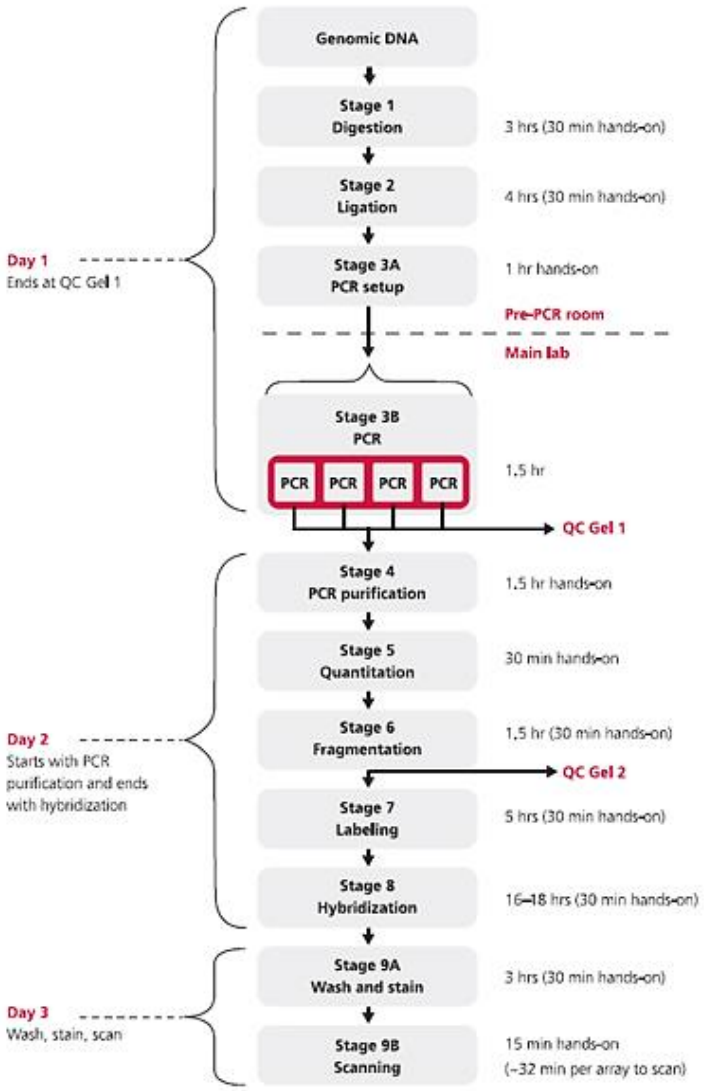

Figure 2.1. Overview of the CytoScan HD protocol (adapted from CytoScan Assay User Manual).
The various stages for the preparation of array CGH. DNA is digested, ligated to adapters then amplified by PCR. The products are purified, fragmented and labelled before overnight hybridization on an array. A wash and stain process prepares the array for scanning. Dates reference the minimum time required for this protocol.

Briefly, 250 ng of genomic DNA from peripheral blood was digested with restriction enzyme *NspI* and ligated to an adapter, followed by PCR amplification in 4 tubes with primers that specifically anneal to the adapter sequence. The PCR products were controlled on a 2% agarose gel in TBE buffer; the expected size ranges between 150 and 2000 bp in length. The tubes from each sample were pooled and purified with magnetic beads to remove contaminants from the PCR reaction. The purified DNA was fragmented using DNase I and visualized on a 4% agarose gel in TBE buffer. This step is the most critical of all and the fragments must be within 25 and 125 bp. The fragments were then labelled with biotin and hybridized overnight at 55ºC and 60 rpm to a CytoScan GeneChip array. The arrays were washed and stained on a GeneChip Fluidics Station 450, and scanned with a GeneChip Scanner 3000 7G, both from Affymetrix.

Raw data files were generated using the software GeneChip Command Console. The data was analyzed with the Chromosome Analysis Suite software version 3.1 and NetAffy genomic annotation file version 33.1 (GRCh37), both developed by Affymetrix. Alterations were queried in the Database of Genomic Variants (DGV) to check for potential common rearrangements, and genes inside or flanking alterations were checked at the OMIM database.

Only CNV of at least 100 kb and with at least 30 supporting probes were considered of interest, as to avoid high level of background noise and false positives for smaller variations with a small number of supporting probes. Nonetheless, CNV of at least 50 kb were checked to confirm whether they affected OMIM genes reported to be related to the proband's clinical phenotype.

2.5. Expression Array

Gene expression in the proband was analyzed using Affymetrix Human Transcriptome 2.0 (HTA 2.0) GeneChip array. This expression array is considered one of the most accurate and extensive tool in the detection of currently known transcript isoforms produced by human genes.

The protocol was implemented in the research group and performed in accordance to the manufacturer's guidelines (GeneChip WT Plus Reagent Kit Manual Target Preparation for GeneChip Whole Transcript Expression Arrays User Manual, http://tools.thermofisher.com/content/sfs/manuals/wtplus_reagentkit_assay_manual.pdf). Figure 2.2. summarizes the procedure, up until the hybridization of samples to the Genechip arrays.

Briefly, 100 ng of total RNA extracted from LCL established from the proband was processed with reagents from WT PLUS reagent kit (Affymetrix) as indicated by manufacturer's protocol. Total RNA was first mixed with Poly-A RNA positive controls. A series of synthesis and purifications are performed, starting with the input RNA until sense-strand cDNA is obtained by the reverse transcription. This cDNA was labelled before an overnight hybridization at 45ºC to a Human Transcriptome 2.0 GeneChip array (Affymetrix). The end goal is to prime the entire length of RNA to provide complete and unbiased coverage of the transcriptome. Hybridized arrays were washed and stained then scanned following the manufacturer's instructions, using GeneChip Fluidics Station FS450 and GCS3000 7G scanner, respectively.

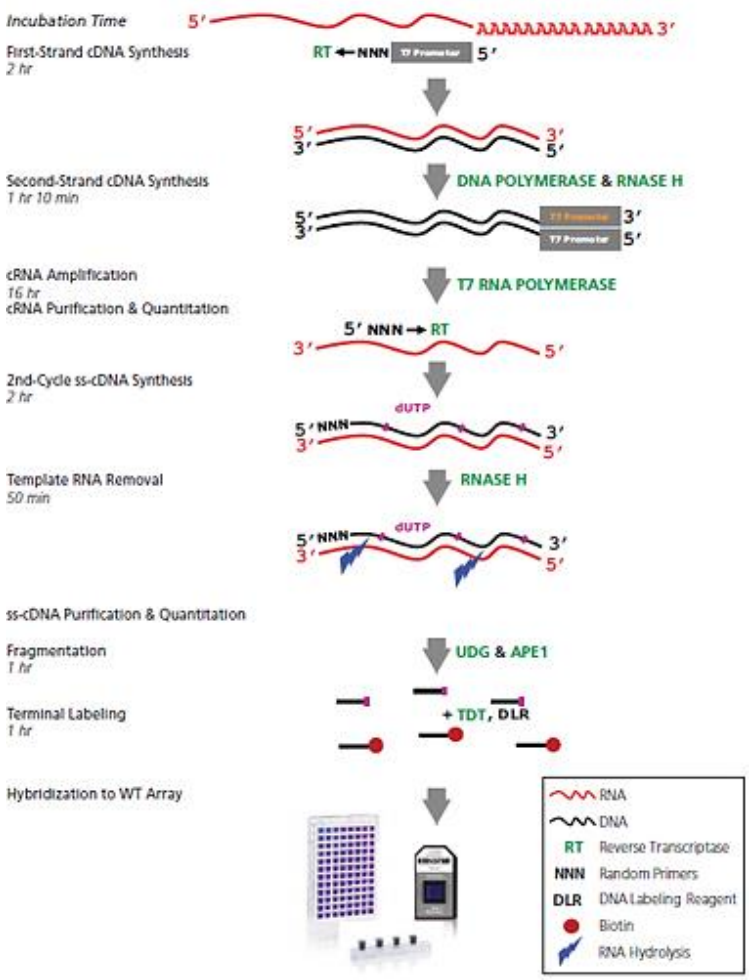

Figure 2.2. Overview of the Affymetrix Human Transcriptome Assay 2.0 protocol (adapted from GeneChip WT PLUS Reagent Kit user guide).
The various stages for the preparation of gene expression array. Synthesis of cDNA from total RNA, followed by cRNA amplification and purification. A second round of cDNA is synthesized, purified and quantified. It is then fragmented and labelled before overnight hybridization on an array. Wash, stain and scanning processes are not here pictured.

The raw data generated were processed in Affymetrix Expression Console software using the RMA algorithm for normalization, background correction and signal summarization. Afterwards, the software Affymetrix Transcriptome Analysis Console was used to identify and compare gene expression levels.

The control group consisted of RNA samples from LCL of unrelated individuals without any of the alterations found in the proband. The control group was analyzed with Affymetrix HTA 2.0, following the protocol described above. Additionally, a set of 12 control individuals considered of normal phenotype on Affymetrix Human Gene 1.0 ST expression array was obtained from a publicly available GEO dataset (GSE42816, http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc= GSE42816). This would give a broader on how variable the expression of certain genes are, independent of the array type used, and also to minimize the risk of using a relatively small HTA 2.0 control group.

2.6. Web Resources

Publicly available databases and online resources on genomic data were frequently queried during the development of this thesis. The web resources are the following:

UCSC Genome Browser (http://genome.ucsc.edu) and *Ensembl* (http://www.ensembl.org) offered general and also positional information on genetic elements of interest (e.g. genes, regulatory elements, repeated regions, SD);

NCBI's dbVar (http://www.ncbi.nlm.nih.gov/dbvar) and the *Database of Genomic Variants* (DGV, http://dgv.tcag.ca/dgv/app/home) were used to check for reported structural variants;

InFest (http://invfestdb.uab.cat) served as hub of reported chromosomal inversion polymorphisms;
Gene Card (http://www.genecards.org) was used for information on genes of interest, namely of those flanking breakpoints;

Mouse Genome Informatics (http://www.informatics.jax.org) and the *Genotype-Tissue Expression Project* (http://www.gtexportal.org) were used for gene expression data in mouse models;

Database of Genomic Variation and Phenotype in Humans using Ensembl Resources (DECIPHER; http://decipher.sanger.ac.uk/) aided the interpretation of genomic variants;

Online Mendelian Inheritance in Man (OMIM, http://www.omim.org) was central for identifying genes with possible clinical significance, and was the main resource for phenotype references.

3. RESULTS AND DISCUSSION

3.1. Clinical report

The proband is male, child of non-consanguineous parents. The parents are both healthy and apparently phenotypically normal. A paternal uncle died at age 5 of unknown causes.

The pregnancy underwent without any complications. However, in the 3rd trimester ultrasound, evidence of delayed intrauterine development was found. Delivery occurred at 41st week of gestation. At birth the proband showed very low birth weight (P5) and length (<P5). Generalized hypertonia and feeding difficulties were described.

Severe congenital malformation syndromes became more apparent as the proband grew up.

At 2 years of age, cryptorchidism with hydrocele was corrected. He took his first steps when he was 3 years old. At 7 years old, the proband presented severe global psychomotor developmental delay and congenital microcephaly was reported. He has facial dysmorphism, with a triangular face, hollow eyes and thin upper lip. He was hospitalized due to intestinal volvulus in the same year.

The timing of basic competences acquisition was very late. He did not develop speech but is able to understand very simple orders and sentences. Autism spectrum disorder (ASD) was diagnosed at infancy, and he is being medicated ever since. He shows stereotypical ASD features such as hyperactivity, attention deficit and can become easily impatient. He has mouthing habits, often bringing foreign objects into his mouth and swallowing them. Feeding difficulties continued through childhood. He suffers from severe chronic constipation, being medicated daily.

Atrial septal defects with interatrial communication (ostium secundum type) was detected and later surgically corrected at 9 years of age. At 15 years old, he still has generalized muscle hypertonia with symmetric myotonic hyperreflexia, being able to mobilize all four limbs without asymmetry. Wide-based gait is noticeable. He has scoliosis with left-side concavity. He has bilateral nail dysplasia and bilateral overlapping of the hallux over the second toe.

Craniocerebral magnetic resonance done during the first year of age showed no relevant alterations. Metabolic analysis performed when he was 15 years old reported normal results for his gender and age group.

Currently, at 16 years of age, he is of short stature (p10) and of very low weight (p<5). He remains completely dependent of aid in all activities of daily life.

3.2. Cytogenetic studies

The pedigree of proband's family is shown in Figure 3.1..

Classical cytogenetic analysis of GLT-banded chromosomes revealed an apparently balanced pericentric inversion at the chromosome 2 in the proband (III:1). More precisely, the inversion

breakpoints were found to be situated at cytobands 2p21 and 2q21. The total number of chromosomes per cell was normal. The subject's karyotype was described as 46, XY, inv(2)(p21q21.1)mat.

Cytogenetic studies were also performed on family members, namely his father and mother (II:3 and II:4). In the mother, a pericentric chromosome 2 inversion, inv(2)(p21q21.1) was found, suggesting that the inversion at chromosome 2 in the proband is of maternal origin. The father harbors a pericentric chromosome 7 inversion, inv(7)(p12.2q21.12), inherited from his mother.

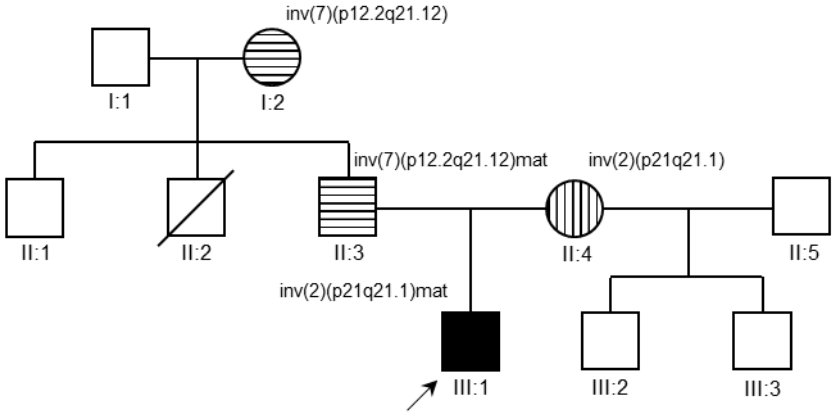

Figure 3.1. Pedigree of the proband's family.
The proband with the inv(2)(p21q21.1)mat, associated with severe congenital malformation syndromes and severe global psychomotor developmental delay, is depicted by a black square (■) and indicated by an arrow. The mother, carrier of an inv(2)(p21q21.1), is depicted by a circle with vertical lines (◍). The proband's father and paternal grandmother, both carriers of an inv(7)(p12.2q21.12), are depicted by horizontal lines filling a square (▤) and a circle (◓), respectively. The paternal uncle (▨) died at childhood of unknown reasons.

3.3. Imbalanced structural alterations

3.3.1. Identification of imbalanced variations

For the screening of imbalanced structural alterations in proband, high resolution genomic array Affymetrix's CytoScan HD was performed. Genomic DNA extracted from whole blood was used.

The genomic array detected several copy number variation in the proband. Most of these structural alterations detected in the proband were small (<100 kb) and located in intergenic regions. The identified variations that directly affected at least one gene often included no OMIM gene, have no phenotypical data available or were not described as associated with the observed clinical features.

For the proband, chromosome structural variations of at least 100 kb in length and with at least 1 reported OMIM gene are shown in Table 3.1.. In particular, a 590 kb duplication was found in 2q21.1, encompassing 13 protein-coding genes, 6 of which described in the OMIM database. The microarray nomenclature is arr[GRCh37] 2q21.1(130,569,840-131,159,016)x4.

Table 3.1. Major structural genomic imbalances in proband

Microarray Nomenclature	Type	Genes	Size (bp)	OMIM Genes	Start Marker	End Marker
arr[GRCh37] 2q21.1(130,569,840-131,159,016)x4	Gain	*LOC389033, LOC100131320, RAB6C, LOC440905, POTEF, CCDC74B-AS1, CCDC74B, SMPD4, MZT2B, TUBA3E, CCDC115, IMP4, PTPN18*	589,176	*RAB6C* (612909), *SMPD4* (610457), *MZT2B* (613450), *CCDC115* (613734), *IMP4* (612981), *PTPN18* (606587)	C-7FBQG	C-3ZABJ
arr[GRCh37] 15q11.2(24,343,759-24,485,858)x1	Loss	*PWRN2*	142,099	*PWRN2* (611217)	C-7CVKO	C-4FRDL
arr[GRCh37] 17q21.31(44,187,491-44,292,319)x3	Gain	*KANSL1, KANSL1-AS1*	104,828	*KANSL1* (612452)	S-3UTFV	S-3FSKZ
arr[GRCh37] Yq11.23(27,014,284-27,226,136)x0	Loss	*DAZ4, DAZ3, DAZ2, BPY2B, BPY2C, BPY2, TTTY4, TTTY4C, TTTY4B*	211,852	*DAZ3* (400027), *DAZ2* (400026)	C-3YBWT	C-7OLRP

[a] Numbers indicate the position of probes in the human genome assembly GRCh37.

Since a possibly pathogenic large duplication is present in the proband, the parents were also analyzed by array CGH Affymetrix's CytoScan HD, in order to understand the origin of the CNV.

In the father, several small sized variations were also identified, most of them not affecting OMIM genes. Notably, a 610 kb duplication was found in chromosome 2 at q21.1. The duplication identified in the father is summarized in Table 3.2.

Table 3.2. The 610 kb duplication in proband's father

Microarray Nomenclature	Type	Genes	Size (bp)	OMIM Genes	Start Marker	End Marker
arr[GRCh37] 2q21.1(130,575,123-131,184,830)x4	Gain	*LOC389033, LOC100131320, RAB6C, LOC440905, POTEF, CCDC74B-AS1, CCDC74B, SMPD4, MZT2B, TUBA3E, CCDC115, IMP4, PTPN18, LOC100216479*	609,707	*RAB6C* (612909), *SMPD4* (610457), *MZT2B* (613450), *CCDC115* (613734), *IMP4* (612981), *PTPN18* (606587)	S-3KEPB	C-3HSSV

[a] Numbers indicate the position of probes in the human genome assembly GRCh37.

The vast majority of the imbalanced structural variations found in the mother are small and do not include any OMIM gene, thus unlikely to be pathogenic.

The 590 kb duplication is further studied and described in the next section.

3.3.2. Chromosome 2 duplication

3.3.2.1. Characterization of the duplication

Array CGH data from CytoScan HD revealed the presence of a large copy number gain in the long arm of chromosome 2, at q21.1. It is 589176 bp in length, located at genomic position chr2:130,569,840-131,159,016 (GRCh37) (Figure 3.2.), between starting marker C-7FBQG and end marker C-3ZABJ. The preceding marker is C-6VUDT and the following marker is C-6EDBO, located in g.130,569,860 and g.131,189,564, respectively. The mean distance between markers in this genomic region is 1575 bp.

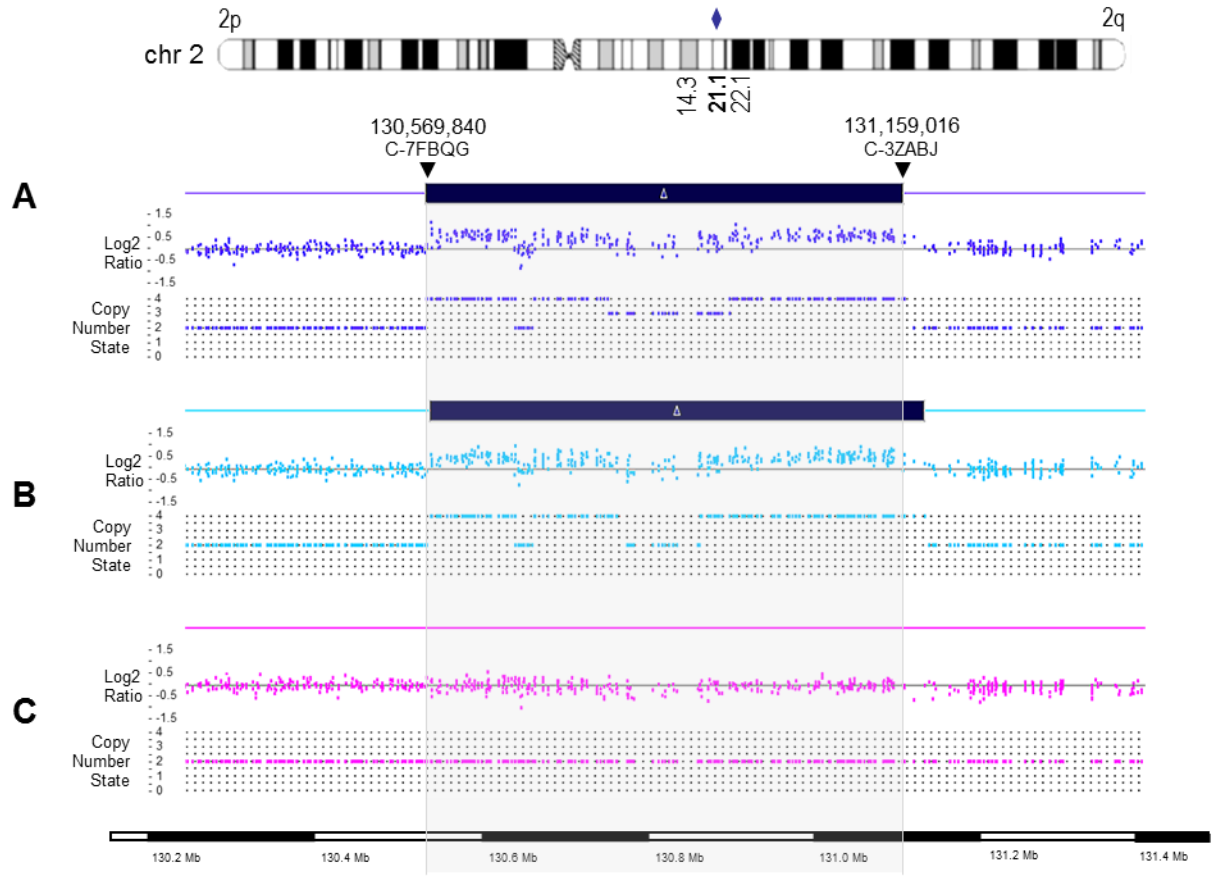

Figure 3.2. Comparative analysis of the 590 kb duplicated genomic region at 2q21.1.
A) The 590 kb duplication at 2q21.1 in the proband. The blue box highlights the duplicated region, described as arr[GRCh37] 2q21.1(130,569,840-131,159,016)x4.
B) The same region from the proband's father. A 610 kb duplication was found, described as arr[GRCh37] 2q21.1(130,575,123-131,184,830)x4.
C) The same region from the proband's mother. No variation in copy number state identified.
Genomic positions are in genome reference assembly GRCh37.

The 610 kb copy number gain identified in proband's father at 2q21.1 is shown in Figure 3.2.B. The 2p21.1 duplication in father and son are very similar in regard to both location and size, differing only by a few markers in the extremities. Since the duplication in the father was interpreted by the software as being slightly bigger, it encompasses 14 genes. The pattern of copy number state variation within the duplication is also very similar between the two. No alteration in this region was detected in proband's mother.

The duplication is in a genomic region with multiple segmental duplications, copy number variations and repeats described in the literature (Figure 3.3). High similarity SD (> 96%) in the region are indicated in the figure. A cluster of SD is located inside the duplicated region, which is also flanked by another cluster at 3'. It is likely that they are linked to the presence of the duplication, since SD have long been considered rearrangements hotspots. However, it is unclear whether the segmental duplications have any direct effect on the proband's phenotype.

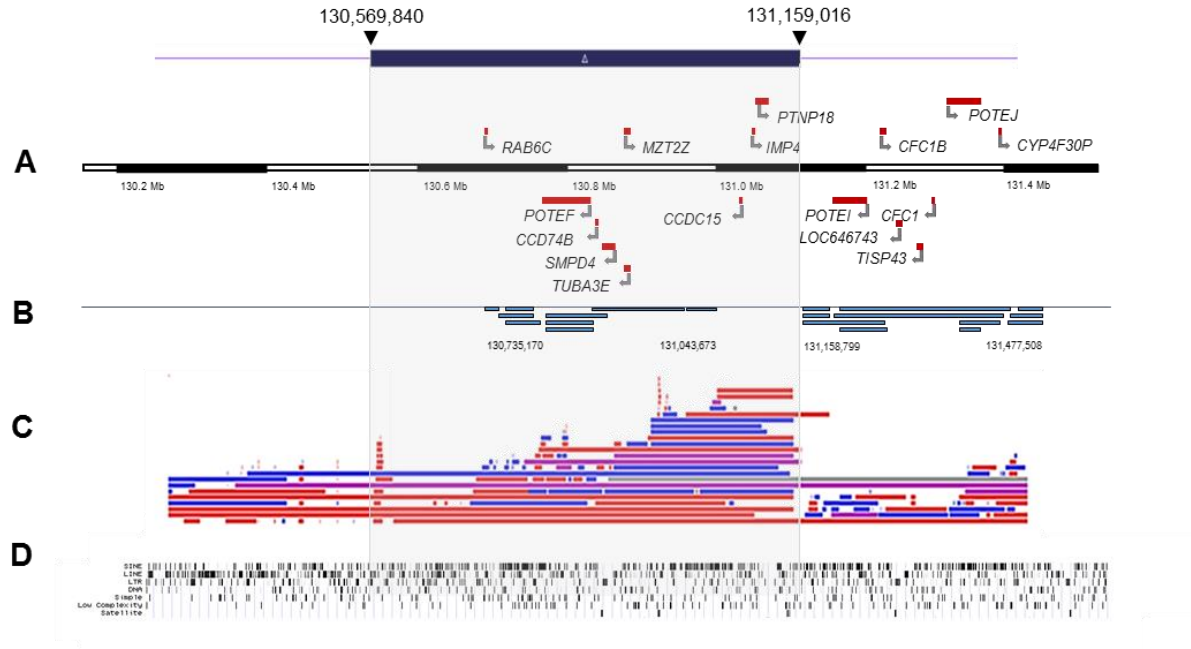

Figure 3.3. Genes, imbalanced variations and repeats in the duplicated genomic region at 2q21.1.
A) Physical map across the duplicated region. Horizontal lines with folded grey arrows indicate the protein-coding genes in sense (above the map) and antisense (below the map) orientation.
B) Clusters of segmental duplications in the region reported in the DGV database. Each blue bar is a segmental duplication with sequence identity between duplications of 96% or greater. The two clusters of segmental duplications are within the genomic locations indicated below (GRCh37).
C) Clusters of CNV reported in the DGV database (blue are duplications, red are deletions, purple are regions of both gain and loss, grey are known variations of uncertain copy number state).
D) Repeating elements in the duplicated region, such as interspersed repeats and low complexity DNA sequences, reported in UCSC Genome Browser.

Variants in copy number are frequently described in this genomic region, be they of gain or loss. The DGV database reports 3 duplications of similar size that overlaps with the 590 kb duplication under study. A 412 kb duplication, located at chr2:130,837,924-131,250,371 (GRCh37) has a 10% overlap (Suktitipat et al. 2014). The 480 kb duplication described by Pinto et al. (2007) has 38% overlap and is located at genome position chr2:130,363,731-131,143,136 (GRCh37). Finally, a smaller duplication of 267 kb, located chr2:130,550,251-130,817,348 (GRCh37), has 61% overlap (Coe et al. 2014). These variations are all from healthy control samples.

Not indicated in DGV, the 374kb duplication reported by Schilter et al. (2013) is located at ch2:130,783,696-131,157,859 (GRCh37) with a 36% overlap with the 590 kb duplication. It is of uncertain pathogenic significance and is occasionally seen in control populations without detectable clinical phenotype.

The presence of these CNV nearby the 590 kb duplication suggest that copy number gains in the region are probably neutral or at least sub-clinical in nature, if considering the variation alone.

The fluctuation of copy number state seen in Figure 3.2. was likely exacerbated by noise related to artifacts in software interpretation of probe signals, possibly influenced by it being a SD rich region (Curtis et al. 2009). Additionally, the pattern could also be partially influenced by the distribution of repeated sequences, such as LINE and SINE, which could further exacerbate the variation of copy number state interpreted by the software. Therefore, the duplication was regarded as having an overall copy number state of x4, though this pattern was taken into consideration when analyzing any results related to the alteration, for example, in gene expression studies.

3.3.2.2. Identification and characterization of genes from the duplicated region

The duplication in the proband encompasses 13 protein-coding genes, 6 of which are found in the OMIM database (Table 3.3). Most of these genes are oncogenes or are of unknown functions.

Table 3.3. Genes within the chromosome 2 duplication

Gene	Gene name	Location[a]	OMIM #
RAB6C	RAB6C, Member RAS Oncogene Family	2:130,737,235-130,740,311	612909
POTEF	POTE Ankyrin Domain Family, Member F	2:130,831,108-130,886,795	-
CCDC74B	Coiled-Coil Domain Containing 74B	2:130,896,860-130,902,707	-
SMPD4	Sphingomyelin Phosphodiesterase 4, Neutral Membrane	2:130,908,981-130,940,323	610457
MZT2B	Mitotic Spindle Organizing Protein 2B	2:130,939,310-130,948,302	613450
TUBA3E	Tubulin, Alpha 3E	2:130,949,318-130,956,034	-
CCDC115	Coiled-Coil Domain Containing 115	2:131,095,814-131,099,922	613734
IMP4	IMP4, U3 Small Nucleolar Ribonucleoprotein, Homolog (Yeast)	2:131,099,798-131,105,383	612981
PTPN18	Protein Tyrosine Phosphatase, Non-Receptor Type 18 (Brain-Derived)	2:131,113,580-131,132,982	606587

[a] Numbers indicate the position of genes in the human genome assembly GRCh37.

The only gene with clinical phenotype associated in OMIM is the Coiled-Coil Domain-Containing Protein 115 (*CCDC115*, chr2:131,095,814-131,099,922, GRCh37; OMIM *613734*)*, also known as Coiled-Coil Protein 1 (*CCP1*). An autosomal recessive congenital disorder of glycosylation (CDG2O;

OMIM #616828) is associated with alteration in the gene (Jasen et al. 2016). It does not appear to be related to the proband's observed phenotype. In mouse, *CCDC115* is mostly expressed in the heart, liver, kidney, and testis, with lower expression in brain and lungs (Pellicano et al. 2006).

Recently, another gene from this duplication was reported to result in pathology. The gene Tubulin Alpha 3E (*TUBA3E*; chr2:130,949,318-130,956,034, GRCh37), yet to be included in OMIM database, is a member of the tubulin family. Alterations in tubulin give rise to tubulinopathies, leading to complex cortical development malformations of varying severity (Bahi-Buisson et al. 2014). A case of tubulinopathy was reported as due to a punctual mutation in *TUBA3E*, with clinical features that includes the autosomal dominant microlissencephaly and global developmental delay (Alazami et al. 2015). Since the cerebral magnetic resonance results found no notable alterations, the proband is unlikely to have microlissencephaly, and thus the contribution of this gene is most likely to be of little significance.

3.3.2.3. Gene expression studies for the duplication

Expression data from the transcriptome assay HTA 2.0 and HuGene 1.0 ST are shown in Table 3.4., for proband and control group known to not harbor the duplication.

Table 3.4. Expression levels of the genes within the chromosome 2 duplication.

Gene	Proband	C1	C2	C3	C4	C5	Control		HuGene 1.0 ST	
							Mean	SD	Mean	SD
LOC389033	3.46	4.92	4.91	5.04	4.88	3.46	4.64	0.66	-	-
RAB6C	3.16	4.15	4.71	4.71	4.34	3.16	4.21	0.64	-	-
POTEF	3.69	4.83	5.4	5.35	5.67	3.69	4.99	0.79	-	-
CCDC74B	4.03	5.31	5.45	5.55	6.25	4.03	5.32	0.81	-	-
SMPD4	7.76	9.32	8.35	8.22	8.93	7.73	8.51	0.62	7.25	0.20
MZT2B	6.04	7.13	6.77	6.78	7.12	6.04	6.77	0.44	5.81	0.22
TUBA3E	5.11	6.22	6.11	6.27	6.23	5.11	5.99	0.49	3.71	0.23
CCDC115	4.87	5.78	5.42	5.64	5.76	4.62	5.44	0.48	6.81	0.19
IMP4	9.19	9.16	8.32	8.81	8.83	8.87	8.80	0.30	9.35	0.24
PTPN18	4.76	6.12	6.19	6.07	6.37	4.76	5.90	0.65	5.65	0.32

Numbers indicate probe signal intensity

The genes in the duplicated region do not appear to be differentially expressed when compared to the control group. Overall, the expression levels of these genes in the proband are relatively similar to the mean value of the control group, despite the existence of the duplication. This may be due to some degree of dosage compensation phenomenon (Kloosterman and Hochstenbach 2014).

The haploinsufficiency index of a given gene denotes the predicted probability of an allele disruption being pathogenic, as it is unable to maintain normal function. A low score (<10%) indicates a high probability of a gene exhibiting haploinsufficiency, while a very low score (>90%) indicates genes likely to not exhibit haploinsufficiency (Huang et al. 2010). The scores of the genes within the

duplicated region is generally low, in average 77.5% with a minimum of 31% as indicated in Decipher database, suggesting that these genes are unlikely to be highly sensitive to copy number changes.

It is of note that, because gene expression analysis was performed on total RNA extracted from LCL, the expression levels detected should be interpreted with a certain degree of caution, due to tissue-specific mechanism in different cell types (Sie et al. 2009). In other words, the expression levels observed in LCL may be different to that of other tissue for some of the genes, even if biological samples from the same person and collected under the same conditions. Nonetheless, LCL are the most widely used tissue for these kind of studies, and was inclusively used in numerous extensive collections like the HapMap (Puig et al. 2015b). Hence, the expression results obtained here are sound for aiding the determination of potential candidate genes.

There are several typical strategies for interpreting pathogenic or benign status for CNV (Miller et al. 2010). For example, if a CNV is identical to one found in an apparently healthy parent or if no OMIM genes were linked to the phenotype, then it is likely a benign variation.

The size and location of the 590 kb duplication in proband suggest that it is most likely the same variation as the 610 kb duplication in his father. It would mean that this duplication was paternally inherited. Since the father was considered phenotypically normal, this may indicate that the duplication is most likely non-pathological if considered as a single isolated alteration.

As such, considering the available data, the 590 kb duplication by itself should be unrelated to the congenital malformation syndrome in the proband.

Since imbalanced alterations were excluded as origin of the observed phenotype, whole-genome sequencing of the proband was performed in order to identify the chromosome 2 inversion breakpoints.

3.4. NGS library preparation and data analysis

In order to determine the chromosome 2 inversion breakpoints, as well as to identify other balanced variations that may have been missed by previously used methods, large-insert whole-genome sequencing (liWGS) was performed. The large-insert libraries were prepared at Talkowski Laboratory at Harvard Medical School, following the protocol described in Talkowski et al. (2011). The library fragments were sequenced on a HiSeq 2000 sequencer (Illumina). The sequencing raw data was analyzed at Talkowski Laboratory and then in-house by the research group.

No outstanding structural alteration aside from the already known chromosome 2 inversion was detected in the proband by NGS.

3.5. Characterization of chromosome 2 inversion

3.5.1. Identification of the inversion breakpoints by NGS

According to the NGS data, the chromosome 2 inversion breakpoints are located in 2p16.1 and in 2q14.3 (Figure 3.4.). The inv2p16.1 breakpoint was localized between genomic positions g.55,934,646-55,935,223 (GRCh37) and the inv2q14.3 breakpoint between g.123,767,565-123,768,077 (GRCh37).

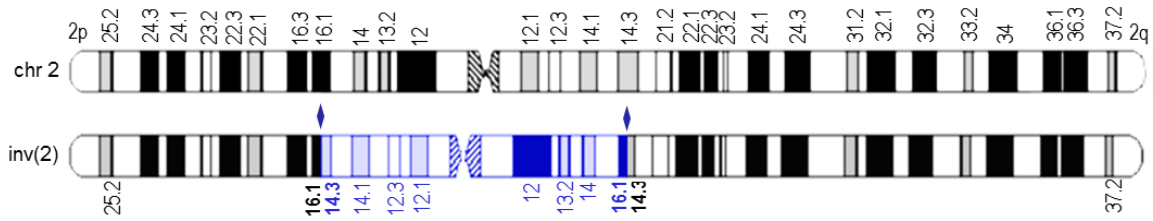

Figure 3.4. Chromosome 2 ideograms.
Ideograms of the wild-type and inv(2)(p16.1;q14.3) chromosomes. Inverted region is in blue, its breakpoints marked with a diamond.

The breakpoint region at inv2p16.1 is supported by 14 read pairs, while inv2q14.3 was supported by 12 read pairs (Figure 3.5.). Although none of the NGS reads mapped exactly on the inversion breakpoints, this approach allowed for a refining of their possible genomic locations into a small interval, delimited by pair reads that map across each of the inversion breakpoints. Consequently, it was possible to amplify these breakpoint regions by conventional PCR based on this information. Afterwards, identification of the exact breakpoints genomic location with a nucleotide resolution was done using Sanger sequencing.

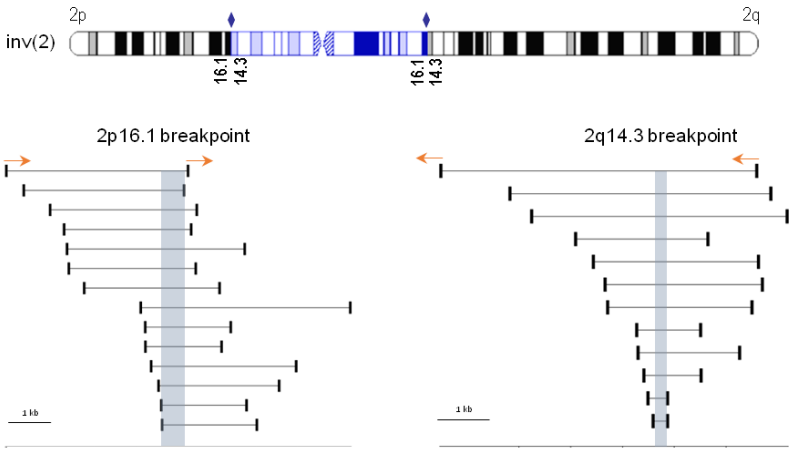

Figure 3.5. NGS read pairs delimiting the inversion breakpoints.
Read pairs delimiting the inversion breakpoints at inv2p16.1 and inv2q14.3, with possible location of breakpoint shaded blue. Black boxes are the reads, and the red arrows indicate their orientation compared to reference genome.

3.5.2. Amplification of inversion junction fragments

The PCR amplification of the inversion breakpoint junction fragments and control fragments at inv2p16.1 and inv2q14.3 is shown in Figure 3.6.. As expected from a maternally inherited alteration, the junction fragments bands in proband and mother are of the same size, 865 bp for the inv2p16.1 fragment and 640 bp for the inv2q14.3 fragment. The control fragments are sized 862 bp and 645 bp, respectively, for the short and long arm breakpoint. The father lacks the alteration, given the observed negative result, as expected.

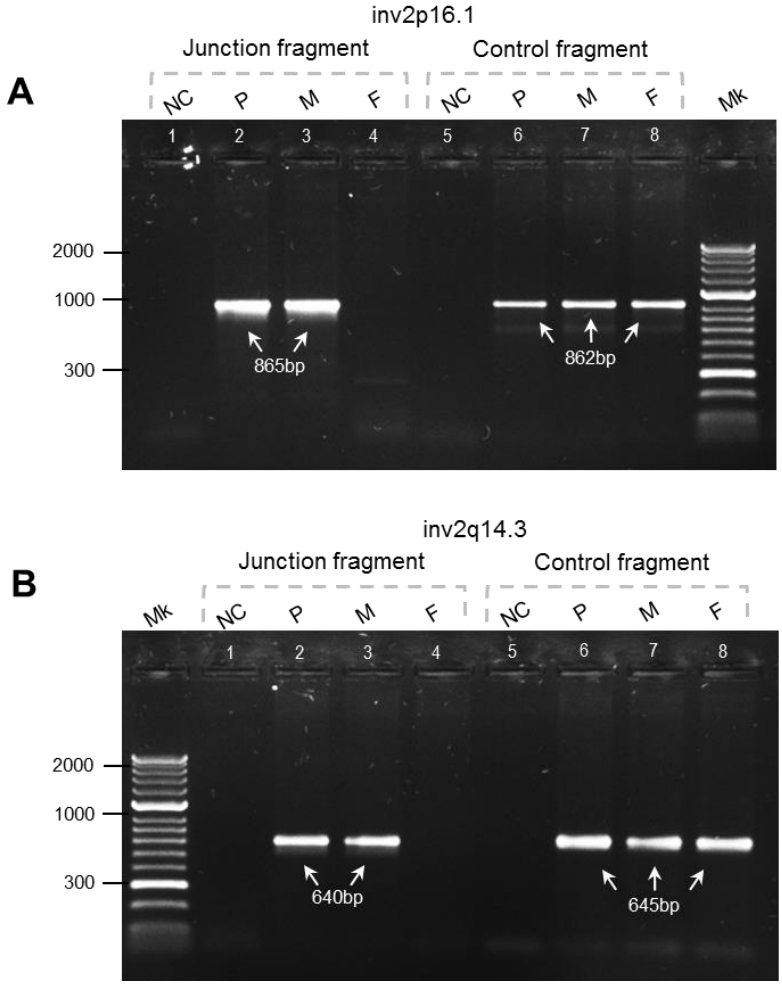

Figure 3.6. Amplification of chromosome 2 inversion junction and control fragments, for proband and parents, on agarose gel electrophoresis.
A) Amplification of the inv2p16.1 junction and control fragments. A 865 bp junction fragment (lanes 1-4) and a 862 bp control fragment (lanes 5-8) were obtained for proband and mother.
B) Amplification of the inv2q14.3 junction and control fragments. A 640 bp junction fragment (lanes 10-13) and a 645 bp control fragment (lanes 14-17) were obtained for proband and mother.
(Mk - HyperLadder 50 bp DNA marker; NC - negative control; P – proband; M – proband's mother; F – proband's father).

Junction fragment amplicons from both proband and his mother were then sequenced by Sanger sequencing, in order to identify the inversion breakpoints with nucleotide resolution. Analysis of the junction fragments determined the inversion breakpoints exact genomic location. Furthermore, the

breakpoints were found to be the same in both proband and his mother, and no nucleotide differences were identified.

The inversion breakpoints are located at position g.55,935,064 (GRCh37) and g.123,767,685 (GRCh37) (Figure 3.7.). By convention, the reported breakpoint was defined as being the last nucleotide able to be read in 5'-3' orientation of the junction fragment, reason why the 5bp deletion was shown as belonging to inv2q14.3 in Figure 3.7.. In term of mechanism, the deleted sequence could have been removed by the repair mechanism after the breakage of the chromosome.

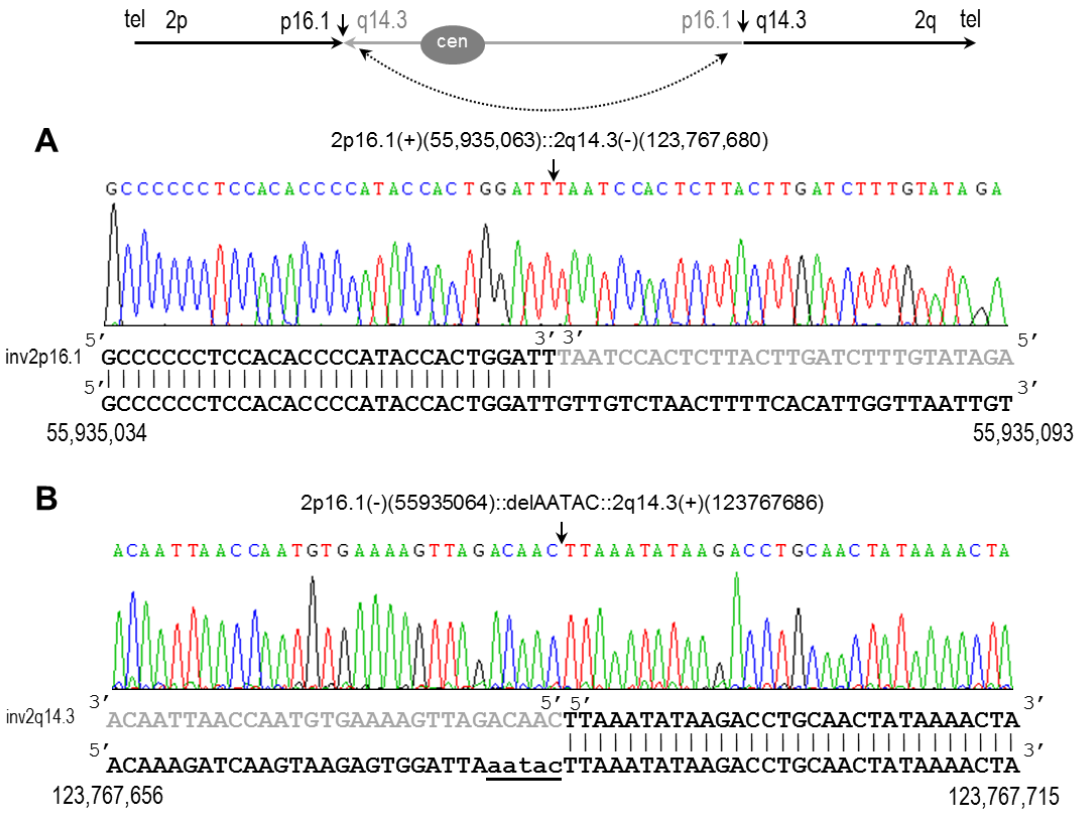

Figure 3.7. Nucleotide sequences of inversion junction fragments aligned against the reference genome sequence.
A) Sequence alignment of inv2p16.1 junction fragment.
B) Sequence alignment of inv2q14.3 junction fragment. A deletion of 5 bases was found, indicated underlined in lower case. On top, the schematic representation of the chromosome 2 inversion, indicating the breakpoint locations. Sequences inside the inversion are in grey, while outside are in black. Junction fragments are shown below the electropherogram. Wild-type chromosome reference sequence at bottom, with their genomic position indicated below. Vertical lines indicate identical nucleotides between inverted and wild-type chromosome. Breakpoints are indicated with a black vertical arrow. Reference genome GRCh37.

The next-generation cytogenetics and sequence-based nomenclature of the inv(2)(p16.1;q14.3) is seq[GRCh37] inv(2)(pter→2p16.1(55,935,06{1-3})::2q14.3(123,767,68{3-1})→2p16.1(55,935,06{5-4})::2q14.3(123,767,68{4-5})→qter) (Ordulu et al. 2014, http://boston.bwh.harvard.edu/input.html). The numbers between the braces reflect the uncertainty of the exact breakpoint.

It is known that segmental duplications can act as hotpots for rearrangements. In the case of the inversion, there are no such features reported nearby the breakpoints. However, interspersed elements

repeats, both long (LINE) and short (SINE), were found just a couple hundred bases from the inv2p16.1 breakpoint, on both sides. The inv2q14.3 breakpoint actually falls inside a LINE called L1M2 (chr2:123,762,071-123,768,419, GRCh37). These repeats have been shown to promote instability and the formation of DNA double-strand breaks (Gilling et al. 2006) and could play a role behind the origin of this rearrangement.

The array CGH results of the proband identified no genomic imbalances nor loss of heterozygosity at the inversion breakpoint regions.

This 68 Mb pericentric chromosomal inversion was initially reported to span from 2p21 to 2q21.1 in the proband and his mother. However, data derived from NGS was able to determine the breakpoints location with incomparably higher resolution, and found that they are actually in the neighboring cytobands p16.1 and q14.3, respectively. Sanger sequencing of the junction fragments identified the breakpoints with nucleotide resolution. Therefore, the subject's karyotype was updated and redefined as 46, XY, inv(2)(p16.1q14.3)mat. It is currently unknown whether the inversion in the mother was inherited or *de novo*.

There are currently no pericentric inversion with breakpoints within 5 Mb of those of inv(2)(p16.1q14.3) reported in the dbVar database. Only a small inversion of 32 kb was reported within 1 Mb of the short arm breakpoint, located at ch2:55,129,764-55,161,515 (GRCh37), without phenotypical data available (Kidd et al. 2008).

3.5.3. Identification of possible candidate genes from the inversion breakpoint regions

The inv2p16.1 breakpoint does not interrupt genes. Figure 3.8. shows the physical map of this region.

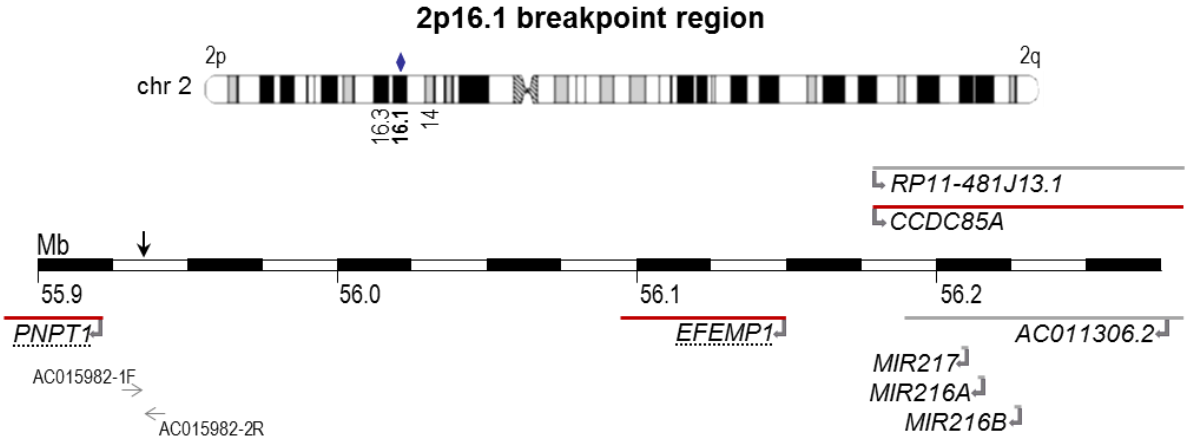

Figure 3.8. Physical map across the inv2p16.1 breakpoint region.
Red horizontal lines with folded arrow indicate the position of genes while grey horizontal lines indicate RNA, in sense (above the map) and antisense (below the map) orientation. Underlined are OMIM genes with phenotypical description available. Black vertical arrow indicate the breakpoint location. Grey horizontal arrow represent primers. Reference genome assembly GRCh37.

The nearest flanking gene is located 14 kb away at 5', the protein-coding Polyribonucleotide Nucleotidyltransferase 1 (*PNPT1*; chr2:55,861,198-55,921,045, GRCh37; OMIM *610316). It codes for a RNA-binding protein, predominantly mitochondrially localized, implicated in numerous RNA metabolic processes (von Ameln et al. 2012). Homozygous mutation in *PNPT1* has been associated with combined oxidative phosphorylation deficiency (COXPD13; OMIM #614932), in which the mutated gene impairs normal import of several RNA species into mitochondria, causing a defect in mitochondrial translation and resulting in mitochondrial respiratory chain deficiency (Vedrenne et al. 2012). The gene is also associated with deafness (DFNB70; OMIM #614934), concordant to the expression studies in mice. Some degree of hearing impairment is typical in patients with a mutated *PNPT1*. Both diseases are autosomal recessive. Since metabolic studies gave normal results and no hearing impairment was detected, it appears that, although so close to the breakpoint, *PNPT1* was not significantly affected by it. In mouse models, this gene is highly expressed in the cochlea and skeletal muscle.

The gene EGF-Containing Fibulin-Like Extracellular Matrix Protein 1 (*EFEMP1*; chr2:56,093,097-56,151,298, GRCh37; OMIM *601548), also known as Fibrillin-Like (*FBNL*), is located about 172 kb from breakpoint at 3'. The extracellular matrix protein coded by *EFEMP1* is very similar to fibrillin, essential for the formation of elastic fibers in connective tissue and may function as a negative regulator of chondrocyte differentiation. It may have an important role in the maintenance of abdominal fascia, as *Efemp1-/-* knockout female mice has pelvic organ support impaired (Rahn et al. 2009). Doyne honeycomb retinal dystrophy (DHRD; OMIM #126600), an autosomal dominant age-related macular degeneration, is caused in a majority of reported cases by mutations in *EFEMP1* (Fu et al. 2007). The proband, however, has no reported ocular defect. In mouse, its highest expression levels are in lung and esophagus, and lowest in heart and brain (Kobayashi et al. 2007).

The inv2q14.3 breakpoint is located in a 'gene-poor' region, far away from protein coding genes, which can be seen in Figure 3.9..

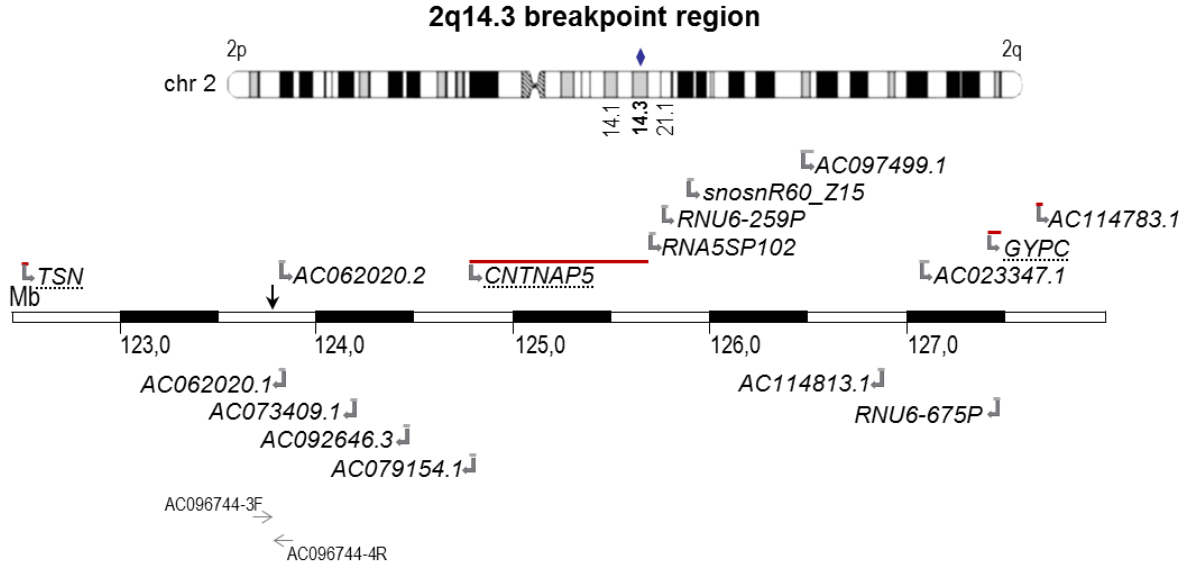

Figure 3.9. Physical map across the inv2q14.3 breakpoint region.
Red horizontal lines indicate the position of genes while grey horizontal lines indicate RNA, in sense (above the map) and antisense (below the map) orientation. Underlined are OMIM genes with phenotypical description available. Black vertical arrow indicate the breakpoint location. Grey horizontal arrow represent primers. Reference genome assembly GRCh37.

Distanced 1.2 Mb proximal is the gene Translin (*TSN*; chr2:122,513,120-122,525,428, GRCh37; OMIM *600575), previously known as Recombination Hotspot-Associated Factor (*RcHF1*). Its coded protein specifically binds to the breakpoint junction translocations found in all acute lymphoblastic leukemia patients studied by Kasai and colleagues (1994). Since it is so connected to translocations, it was promptly renamed Translin (Aoki et al. 1995). *TSN* is involved in some degree with DNA damage repair as well as RNA trafficking in neurons (Li et al. 2008). Regarding human pathologies related to *TSN*, only in leukemia was it reported. The gene is most expressed in mouse subcutaneous adipose tissue. In knockout mice a wide range of behavior alterations, as well as lower reproductive success, were noticed (Stein et al. 2006).

The gene Contactin-Associated Protein-Like 5 (*CNTNAP5*; chr2:124,782,863-125,672,953, GRCh37; OMIM *610519) is the closest protein-coding gene to the inv2q14.3 breakpoint, but is still distanced over 1 Mb distal. The gene *CNTNAP5* produces a protein involved in cell adhesion and intercellular communication. Some studies propose that small cryptic mutations disrupting *CNTNAP5* may be linked to susceptibility to dyslexia and ASD (An et al. 2014; Pagnamenta et al. 2010). It has increased expression in mice's nervous system and most predominantly expressed in subcutaneous adipose cells (Traut et al. 2006).

The remaining protein-coding genes in inv2p16.1 and inv2q14.3 regions were also studied, namely the Coiled-Coil Domain Containing 85A (*CCDC85A*; chr2:56,411,258-56,613,309, GRCh37), Glycophorin C (*GYPC*; chr2:127,413,510-127,454,250, GRCh37; OMIM *110750) and Uncharacterized LOC101929926 (*AC114783* or *LOC101929926*; ch2:127,656,458-127,659,673, CRCh37). However, they have either little information publicly available or have no reports that supported them as possibly involved with the malformation syndrome in proband, either due to their function, expression in mouse or associated clinical phenotypes. For example, mutations in gene *GYPC* are mainly associated with resistance to malaria. Consequently, these genes are likely excluded as being involved in the observed phenotype.

For structural chromosomal anomalies, the prime aspect in determining their significance is where the breakpoints are located and if these interrupt a candidate gene or lead to disruption of their transcriptional regulation (Feuk 2010). This is not the case of this inversion, where the breakpoints junctions did not directly disrupt any gene, and no evidence of known regulatory elements sitting at both coordinates was found.

3.5.4. Gene expression analysis for the inversion

The expression levels obtained from HTA 2.0 and HuGene 1.0 ST of genes flanking the inversion 2 breakpoints, as well as nearby protein-coding genes, are shown in Table 3.5.

The levels in the proband appear to be not significantly different from that of the control group. The standard deviation in these genes are high compared to that from HuGene 1.0 ST, possibly due to higher sensitivity of the newer platform and the small size of the control group.

Table 3.5 Expression levels of genes in the inversion 2 breakpoints regions

Gene	Proband	C1	C2	C3	C4	C5	Control		HuGene 1.0 ST	
							Mean	SD	Mean	SD
inv2p16.1 region										
PNPT1	9.66	10.8	11.05	11.00	10.32	9.64	10.56	0.59	9.29	0.33
EFEMP1	3.88	6.27	4.89	5.01	4.95	3.76	4.98	0.89	3.49	0.13
inv2q14.3 region										
TSN	8.83	10.08	10.17	9.79	9.95	8.83	9.76	0.54	8.59	0.26
CNTNAP5	3.13	4.07	4.16	4.06	4.13	3.13	3.91	0.44	3.41	0.17
GYPC	5.50	8.89	6.63	7.35	7.84	6.46	7.43	0.99	7.29	0.41
AC114783	2.85	3.88	4.48	4.1	3.88	2.85	3,84	0,60	-	-

Numbers indicate probe signal intensity.

Summarizing the results so far, the chromosome 2 inversion is confirmed to be maternal in heritage. The breakpoints do not disrupt genes, with one of the breakpoints located in a gene poor region. The protein-coding genes flanking the inversion were not considerably affected by the rearrangement in terms of their expression levels in LCL. The remaining genes nearby the breakpoints are also unlikely to be related to the phenotype. It appears that existence of this pericentric inversion by itself should not have caused the observed malformation syndrome.

Taking these findings into consideration, the question of whether the inversion is polymorphic in nature, but at a low frequency, is raised.

Additionally, the existence of environmental factors during embryogenesis that may have led to the clinical phenotype in the proband is implausible given the currently available medical data.

4. CONCLUSIONS AND FUTURE WORK

This study aimed to identify the candidate genes and molecular alterations behind a severe malformation syndrome. The case study focuses in an individual with an apparently balanced pericentric inversion defined by cytogenetic analysis as inv(2)(p21;q21.1)mat.

First, imbalanced genomic anomalies were screened in the proband, using genomic arrays. A plausible candidate for pathogenesis was the 590 kb duplication in 2q21.1. A very similar 610 kb duplication was discovered in proband's father, who is phenotypically normal, being supposedly the same alteration. The 590 kb duplication affects several OMIM genes, but only one gene, CCDC115, has been associated with a clinical phenotype to date, but is unrelated to the proband's malformation syndrome. Expression studies found that their levels appear not significantly different from the control group. The region is also reported with several CNV of no or uncertain pathogenic significance, as well as segmental duplications. Taking these findings into account, especially due to its paternal heritage, the 2q21.1 duplication by itself is most likely nonpathogenic.

Concerning the chromosome 2 inversion, long-insert whole genome sequencing was performed to identify the inversion breakpoints as well as any possible cryptic alterations. The results obtained from NGS data analysis provided deeper understanding of the balanced rearrangement in study and expedited the determination of breakpoints. NGS proved to be an invaluable method for the determination of inversion breakpoints, since it was able to identify the chromosome 2 inversion with much higher resolution than with cytogenetic studies. The proband's karyotype was redefined to 46, XY, inv(2)(p16.1q14.3)mat. The breakpoints were determined at nucleotide resolution by sequencing of the junction fragments. Also, familiar segregation study allowed the confirmation that the inversions found in the proband and proband's mother is identical. This result supports the refutation of disease-causing effect of the inversion if considering the rearrangement by itself.

Furthermore, the inversion does not disrupt any gene and the publicly accessible information on breakpoints flanking genes, such as their biological function and associated pathologies, gave little support to possible phenotype-genotype relationship with the proband's clinical features. Additionally, expression studies were undertaken and the data suggest that genes closest to the breakpoints do not have a significantly altered expression level. It is unlikely that these genes alone have a significant role in pathogenesis.

Also, by array CGH, no CNV nor loss of heterozygosity were found in the inversion breakpoints regions, or in their corresponding cytoband.

Considering the data obtained using these methodologies, is unlikely that the two rearrangements here described, alone, could be behind the observed clinical phenotype in the proband, since the parents are phenotypically normal. In the end, the inversion was excluded as the sole explanation of the severe malformation syndrome. Consequently, the question of whether the inversion is polymorphic in nature, but at a low frequency, is raised. It can't be excluded the influence on the congenital malformation of environmental factors during embryogenesis.

At the moment, genetic etiology of the congenital malformation is yet to be determined. Thus, further studies are warranted for the identification of the molecular alterations responsible for the observed

congenital malformation syndrome. Also, expanding the gene expression analysis genome-wide could help elucidate the proband's phenotype.

It is possible that the combination of chromosome 2 inversion and the duplication at 2q21.1 could lead to disease, even if they alone appear insufficient. More studies are necessary to infer whether the phenotype could have been exacerbated by the interaction between of them, or between yet to be identified alterations, like the phenomenon previously described by David et al. (2015).

Similarly, it is possible that the malformation syndrome could be due to cryptic mutations undetected by currently utilized methodologies. Therefore, exome sequencing is proposed for this task. Whole-exome sequencing is traditionally used to detect single-nucleotide variants and small indels. Since exome sequencing focuses in the coding regions with its capture-based approach, it has a very high coverage, increasing the confidence in any detected point mutations. Additionally, it has the advantage of being faster and more economically accessible than whole-genome sequencing (Yang et al. 2016). The downside is the complexity of data analysis, sometimes yielding inconclusive findings. Nonetheless, by sequencing the proband’s exome, this highly complex approach should be able to detect alterations missed by other methods and possibly identify the molecular cause behind the clinical phenotype.

Even though not the object of this thesis, preparation of large-insert whole genome sequencing libraries based on the protocol by Talkowski et al. (2011) and its updated version (Hanscom and Talkowski 2014) was implemented in the research group. This would be of particular importance for large scale application of this approach for identification of structural chromosomal anomalies that will represent a hallmark in the study of chromosome rearrangements associated with congenital malformations.

5. REFERENCES

Aguado C, Gayà-Vidal M, Villatoro S, Oliva M, Izquierdo D, Giner-Delgado C, Montalvo V, García-González J, Martínez-Fundichely A, Capilla L, Ruiz-Herrera A, Estivill X, Puig M, Cáceres M (2014) Validation and genotyping of multiple human polymorphic inversions mediated by inverted repeats reveals a high degree of recurrence. PLoS Genet 10:e1004208

Alazami AM, Patel N, Shamseldin HE, Anazi S, Al-Dosari MS, Alzahrani F, Hijazi H, Alshammari M, Aldahmesh MA, Salih MA, Faqeih E (2015) Accelerating Novel Candidate Gene Discovery in Neurogenetic Disorders via Whole-Exome Sequencing of Prescreened Multiplex Consanguineous Families. Cell reports 10(2):148–161

Alves JM, Lopes AM, Chikhi L, Amorim A (2012) On the structural plasticity of the human genome: chromosomal inversions revisited. Curr Genomics 13:623–32

An JY, Cristino AS, Zhao Q, Edson J, Williams SM, Ravine D, Wray J, Marshall VM, Hunt A, Whitehouse AJ, Claudianos C (2014) Towards a molecular characterization of autism spectrum disorders: an exome sequencing and systems approach. Transl Psychiatry 4:e394

Antonacci F, Kidd JM, Marques-Bonet T, Ventura M, Siswara P, Jiang Z, Eichler EE (2009) Characterization of six human disease-associated inversion polymorphisms. Hum Mol Genet 18:2555–66

Aoki K, Suzuki K, Sugano T, Tasaka T, Nakahara K, Kuge O, Omori A, Kasai M (1995) A novel gene, Translin, encodes a recombination hotspot binding protein associated with chromosomal translocations. Nat Genet 10:167–74

Avelar TA, Perfeito L, Gordo I, Ferreira MG (2013) Genome architecture is a selectable trait that can be maintained by antagonistic pleiotropy. Nat Commun 4:2235

Bahi-Buisson N, Poirier K, Fourniol F, Saillour Y, Valence S, Lebrun N, Hully M, Bianco CF, Boddaert N, Elie C, Lascelles K (2014) The wide spectrum of tubulinopathies: what are the key features for the diagnosis? Brain 137:1676–700

Batista DAS, Pai GS, Stetten G (1994) Molecular analysis of a complex chromosomal rearrangement and a review of familial cases. Am J Med Genet 53:255–263

Bhat TA, Wani AA (2017) Chromosome Structure and Aberrations. Springer India. New Delhi

Bhatt S, Moradkhani K, Mrasek K, Puechberty J, Lefort G, Lespinasse J, Sarda P, Liehr T, Hamamah S, Pellestor F (2007) Breakpoint characterization: a new approach for segregation analysis of paracentric inversion in human sperm. Mol Hum Reprod 13:751–6

Cáceres A, González JR (2015) Following the footprints of polymorphic inversions on SNP data: from detection to association tests. Nucleic Acids Res 43:e53

Cáceres A, Sindi SS, Raphael BJ, Cáceres M, González JR (2012) Identification of polymorphic inversions from genotypes. BMC Bioinformatics 13:28

Chen W, Ullmann R, Langnick C, Menzel C, Wotschofsky Z, Hu H, Döring A, Hu Y, Kang H, Tzschach A, Hoeltzenbein M, Neitzel H, Markus S, Wiedersberg E, Kistner G, van Ravenswaaij-Arts C, Kleefstra T, Kalscheuer V, Ropers HH (2010) Breakpoint analysis of balanced chromosome rearrangements by next-generation paired-end sequencing. Eur J Hum Genet 18:539–543

Coe BP, Witherspoon K, Rosenfeld JA, Van Bon BW, Vulto-van Silfhout AT, Bosco P, Friend KL, Baker C, Buono S, Vissers LE, Schuurs-Hoeijmakers JH (2014) Refining analyses of copy number variation identifies specific genes associated with developmental delay. Nat Genet 46(10):1063–71

Cohen MM, Rosenmann A, Hacham- Zadeh S, Dahan S (1975) Dicentric X- isochromosome (Xqi dic) and pericentric inversion of No. 2 inv (2)(p15) q21) in a patient with gonadal dysgenesis. Clin Genet 8(1):11–7

Corbett-Detig RB, Cardeno C, Langley CH (2012) Sequence-based detection and breakpoint assembly of polymorphic inversions. Genetics 192:131–7

Corsello G, Giuffrè M (2012) Congenital malformations. J Matern Fetal Neonatal Med 25(sup1):25–9

Curtis C, Lynch AG, Dunning MJ, Spiteri I, Marioni JC, Hadfield J, Chin SF, Brenton JD, Tavaré S, Caldas C (2009) The pitfalls of platform comparison: DNA copy number array technologies assessed. BMC Genomics 10:588

Czepulkowski B (2001) Analyzing Chromosomes. BIOS Scientific. Oxford

Dahm R (2008) Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. Hum Genet 122:565–81

David D, Almeida LS, Maggi M, Araújo C, Imreh S, Valentini G, Fekete G, Haltrich I (2015) Clinical Severity of PGK1 Deficiency Due To a Novel p.E120K Substitution Is Exacerbated by Co-inheritance of a Subclinical Translocation t(3;14)(q26.33;q12), Disrupting NUBPL Gene. JIMD Rep 23:55–65

David D, Marques B, Ferreira C, Araújo C, Vieira L, Soares G, Dias C, Pinto M (2013) Co-segregation of trichorhinophalangeal syndrome with a t(8;13)(q23.3;q21.31) familial translocation that appears to increase TRPS1 gene expression. Hum Genet 132:1287–99

David D, Marques B, Ferreira C, Vieira P, Corona-Rivera A, Ferreira JC, van Bokhoven H (2009) Characterization of two ectrodactyly-associated translocation breakpoints separated by 2.5 Mb on chromosome 2q14.1-q14.2. Eur J Hum Genet 17:1024–33

De Gregori M, Ciccone R, Magini P, Pramparo T, Gimelli S, Messa J, Novara F, Vetro A, Rossi E, Maraschio P, Bonaglia MC, Anichini C, Ferrero GB, Silengo M, Fazzi E, Zatterale A, Fischetto R, Previderé C, Belli S, Turci A, Calabrese G, Bernardi F, Meneghelli E, Riegel M, Rocchi M, Guerneri S, Lalatta F, Zelante L, Romano C, Fichera M, Mattina T, Arrigo G, Zollino M, Giglio S, Lonardo F, Bonfante A, Ferlini A, Cifuentes F, Van Esch H, Backx L, Schinzel A, Vermeesch JR, Zuffardi O (2007) Cryptic deletions are a common finding in “balance” reciprocal and complex chromosome rearrangements: a study of 59 patients. J Med Genet 44:750–62

Desjardins P, Conklin D (2010) NanoDrop microvolume quantitation of nucleic acids. JoVE 22(45):e2565-e2565

Djalali M, Steinbach P, Bullerdiek J, Holmes-Siedle M, Verschraegen-Spae MR, Smith A (1986) The significance of pericentric inversions of chromosome 2. Hum gen 72(1):32-6

El-Baz F, Zaghloul MS, El Sobky E, Elhossiny RM, Salah H, Abdelaziz NE (2016) Chromosomal abnormalities and autism. Egypt J Med Hum Genet 17(1):57–62

Entesarian M, Carlsson B, Mansouri MR, Stattin EL, Holmberg E, Golovleva I, Stefansson H, Klar J, Dahl N (2009) A chromosome 10 variant with a 12 Mb inversion [inv(10)(q11.22q21.1)] identical by descent and frequent in the Swedish population. Am J Med Genet A 149A:380–6

Feuk L (2010) Inversion variants in the human genome: role in disease and genome architecture. Genome Med 2:11

Fickelscher I, Liehr T, Watts K, Bryant V, Barber JC, Heidemann S, Siebert R, Hertz JM, Tumer Z, Simon Thomas N (2007) The variant inv(2)(p11.2q13) is a genuinely recurrent rearrangement but displays some breakpoint heterogeneity. Am J Hum Genet 81:847–56

Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, Aburatani H, Jones KW, Tyler-Smith C, Hurles ME, Carter NP, Scherer SW, Lee C (2006) Copy number variation: new insights in genome diversity. Genome Res 16:949–61

Fu L, Garland D, Yang Z, Shukla D, Rajendran A, Pearson E, Stone EM, Zhang K, Pierce EA (2007) The R345W mutation in EFEMP1 is pathogenic and causes AMD-like deposits in mice. Hum Mol Genet 16:2411–22

Genetics Home Reference (GHR) (2014) Help Me Understand Genetics - Handbook

Gilling M, Dullinger JS, Gesk S, Metzke-Heidemann S, Siebert R, Meyer T, Brondum-Nielsen K, Tommerup N, Ropers HH, Tumer Z, Kalscheuer VM, Thomas NS (2006) Breakpoint cloning and haplotype analysis indicate a single origin of the common Inv(10)(p11.2q21.2) mutation among northern Europeans. Am J Hum Genet 78:878–83

Grada A, Weinbrecht K (2013) Next-generation sequencing: methodology and application. J Invest Dermatol 133:e11.

Griffiths A, Miller J, Suzuki D, Lewontin R, Gelbart W (2000) An Introduction to Genetic Analysis, 7th edn. New York

Griffiths AJ, Gelbart WM, Miller JH, Lewontin RC (1999) Modern Genetic Analysis. W. H. Freeman, New York

Hanscom C, Talkowski M (2014) Design of large-insert jumping libraries for structural variant detection using Illumina sequencing. Curr Protoc Hum Genet 80:7.22.1-9

Higgins AW, Alkuraya FS, Bosco AF, Brown KK, Bruns GA, Donovan DJ, Eisenman R, Fan Y, Farra CG, Ferguson HL, Gusella JF (2008) Characterization of apparently balanced

chromosomal rearrangements from the developmental genome anatomy project. Am J Hum Genet 82(3):712–22

Hillmer AM, Yao F, Inaki K, Lee WH, Ariyaratne PN, Teo AS, Woo XY, Zhang Z, Zhao H, Ukil L, Chen JP, Zhu F, So JB, Salto-Tellez M, Poh WT, Zawack KF, Nagarajan N, Gao S, Li G, Kumar V, Lim HP, Sia YY, Chan CS, Leong ST, Neo SC, Choi PS, Thoreau H, Tan PB, Shahab A, Ruan X, Bergh J, Hall P, Cacheux-Rataboul V, Wei CL, Yeoh KG, Sung WK, Bourque G, Liu ET, Ruan Y (2011) Comprehensive long-span paired-end-tag mapping reveals characteristic patterns of structural variations in epithelial cancer genomes. Genome Res 21:665–75

Honeywell C, Argiropoulos B, Douglas S, Blumenthal AL, Allanson J, McGowan-Jordan J, McCready ME (2012) Apparent transmission distortion of a pericentric chromosome one inversion in a large multi-generation pedigree. Am J Med Genet Part A 158A:1262–1268

Huang N, Lee I, Marcotte EM, Hurles ME (2010) Characterising and predicting haploinsufficiency in the human genome. PLoS Genet 6(10):e1001154

Jansen JC, Cirak S, van Scherpenzeel M, Timal S, Reunert J, Rust S, Pérez B, Vicogne D, Krawitz P, Wada Y, Ashikov A (2016) CCDC115 deficiency causes a disorder of Golgi homeostasis with abnormal protein glycosylation. Am J Hum Genet 98(2):310-21

Kasai M, Aoki K, Matsuo Y, Minowada J, Maziarz RT, Strominger JL (1994) Recombination hotspot associated factors specifically recognize novel target sequences at the site of interchromosomal rearrangements in T-ALL patients with t(8;14)(q24;q11) and t(1;14)(p32;q11). Int Immunol 6:1017–25

Kidd JM, Cooper GM, Donahue WF, Hayden HS, Sampas N, Graves T, Hansen N, Teague B, Alkan C, Antonacci F, Haugen E (2008) Mapping and sequencing of structural variation from eight human genomes. Nature 453(7191):56–64

Kirkpatrick M (2010) How and why chromosome inversions evolve. PLoS Biol 8(9): e1000501

Kjeldsen E (2015) A novel acquired inv (2)(p23. 3q24. 3) with concurrent submicroscopic deletions at 2p23. 3, 2p22. 1, 2q24. 3 and 1p13. 2 in a patient with chronic thrombocytopenia and anemia. Mol cytogenet 8(1):7

Kobayashi N, Kostka G, Garbe JH, Keene DR, Bächinger HP, Hanisch FG, Markova D, Tsuda T, Timpl R, Chu ML, Sasaki T (2007) A comparative analysis of the fibulin protein family. Biochemical characterization, binding interactions, and tissue localization. J Biol Chem 282:11805–16

Le Caignec C, Boceno M, Saugier-Veber P, Jacquemont S, Joubert M, David A, Frebourg T, Rival JM (2005) Detection of genomic imbalances by array based comparative genomic hybridisation in fetuses with multiple malformations. J Med Genet 42(2):121–8

Leonard DGB (2016) Molecular Pathology in Clinical Practice, 2nd edition. Springer International Publishing Switzerland. Cham

Li Z, Wu Y, Baraban JM (2008) The Translin/Trax RNA binding complex: Clues to function in the nervous system. Biochim Biophys Acta - Gene Regul Mech 1779:479–485

Liu P, Erez A, Nagamani SC, Dhar SU, Kołodziejska KE, Dharmadhikari AV, Cooper ML, Wiszniewska J, Zhang F, Withers MA, Bacino CA, Campos-Acevedo LD, Delgado MR, Freedenberg D, Garnica A, Grebe TA, Hernández-Almaguer D, Immken L, Lalani SR, McLean SD, Northrup H, Scaglia F, Strathearn L, Trapane P, Kang SH, Patel A, Cheung SW, Hastings PJ, Stankiewicz P, Lupski JR, Bi W (2011) Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. Cell 146:889–903

Luthardt FW, Keitges E (2001) Chromosomal Syndromes and Genetic Disease. In: Encyclopedia of Life Sciences. John Wiley & Sons, Ltd, Chichester, UK

MacIntyre DJ, Blackwood DH, Porteous DJ, Pickard BS, Muir WJ (2003) Chromosomal abnormalities and mental illness. Mol Psychiatry 8(3):275–87

Mardis ER (2008) Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 9:387–402

Martínez-Fundichely A, Casillas S, Egea R, Ràmia M, Barbadilla A, Pantano L, Puig M, Cáceres M (2014) InvFEST, a database integrating information of polymorphic inversions in the human genome. Nucleic Acids Res 42:D1027-32

Metzker ML (2010) Sequencing technologies - the next generation. Nat Rev Genet 11:31–46

Mihelec M, St Heaps L, Flaherty M, Billson F, Rudduck C, Tam PP, Grigg JR, Peters GB, Jamieson RV (2008) Chromosomal rearrangements and novel genes in disorders of eye development, cataract and glaucoma. Twin Res Hum Genet 11:412–21

Miller DT, Adam MP, Aradhya S, Biesecker LG, Brothman AR, Carter NP, Church DM, Crolla JA, Eichler EE, Epstein CJ, Faucett WA, Feuk L, Friedman JM, Hamosh A, Jackson L, Kaminsky EB, Kok K, Krantz ID, Kuhn RM, Lee C, Ostell JM, Rosenberg C, Scherer SW, Spinner NB, Stavropoulos DJ, Tepperberg JH, Thorland EC, Vermeesch JR, Waggoner DJ, Watson MS, Martin CL, Ledbetter DH (2010) Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. Am J Hum Genet 86:749–64

Muss B, Schwanitz G (2007) Characterization of Inversions as a Type of Structural Chromosome Aberration. Int J Hum Genet 7(2):141–161

Nambiar M, Raghavan SC (2011) How does DNA break during chromosomal translocations?. Nucleic Acids Res 39:5813–25

Obe G, Pfeiffer P, Savage JRK, Johannes C, Goedecke W, Jeppesen P, Natarajan AT, Martınez-López W, Folle GA and Drets ME (2002) Chromosomal aberrations: formation, identification and distribution. Elsevier Fundam Mol Mech Mutagen 504:17–36

Ordulu Z, Kammin T, Brand H, Pillalamarri V, Redin CE, Collins RL, Blumenthal I, Hanscom C, Pereira S, Bradley I, Crandall BF, Gerrol P, Hayden MA, Hussain N, Kanengisser-Pines B, Kantarci S, Levy B, Macera MJ, Quintero-Rivera F, Spiegel E, Stevens B, Ulm JE, Warburton D, Wilkins-Haug LE, Yachelevich N, Gusella JF, Talkowski ME, Morton CC (2016) Structural Chromosomal Rearrangements Require Nucleotide-Level Resolution: Lessons from Next-Generation Sequencing in Prenatal Diagnosis. Am J Hum Genet 99(5):1015–1033

Ordulu Z, Wong KE, Currall BB, Ivanov AR, Pereira S, Althari S, Gusella JF, Talkowski ME, Morton CC (2014) Describing sequencing results of structural chromosome rearrangements with a suggested next-generation cytogenetic nomenclature. Am J Hum Genet 94:695–709

Pagnamenta AT, Bacchelli E, de Jonge MV, Mirza G, Scerri TS, Minopoli F, Chiocchetti A, Ludwig KU, Hoffmann P, Paracchini S, Lowy E, Harold DH, Chapman JA, Klauck SM, Poustka F, Houben RH, Staal WG, Ophoff RA, O'Donovan MC, Williams J, Nöthen MM, Schulte-Körne G, Deloukas P, Ragoussis J, Bailey AJ, Maestrini E, Monaco AP; International Molecular Genetic Study Of Autism Consortium (2010) Characterization of a family with rare deletions in CNTNAP5 and DOCK4 suggests novel risk loci for autism and dyslexia. Biol Psychiatry 68:320–8

Paulson JR, Vagnarelli P (2011) Chromosomes and Chromatin. In: Encyclopedia of Life Sciences. John Wiley & Sons, Ltd. Chichester

Pellicano F, Inglis-Broadgate SL, Pante G, Ansorge W, Iwata T (2006) Expression of coiled-coil protein 1, a novel gene downstream of FGF2, in the developing brain. Gene Expr Patterns 6:285–93

Pettersson E, Lundeberg J, Ahmadian A (2009) Generations of sequencing technologies. Genomics 93:105–11

Pinto D, Marshall C, Feuk L, Scherer SW (2007) Copy-number variation in control population cohorts. Hum Mol Genet R168-73

Puig M, Casillas S, Villatoro S, Cáceres M (2015b) Human inversions and their functional consequences. Brief Funct Genomics 14:369–379

Puig M, Castellano D, Pantano L, Giner-Delgado C, Izquierdo D, Gayà-Vidal M, Lucas-Lledó JI, Esko T, Terao C, Matsuda F, Cáceres M (2015a) Functional Impact and Evolution of a Novel Human Polymorphic Inversion That Disrupts a Gene and Creates a Fusion Transcript. PLoS Genet 11:e1005495

Rahn DD, Acevedo JF, Roshanravan S, Keller PW, Davis EC, Marmorstein LY, Word RA (2009) Failure of pelvic organ support in mice deficient in fibulin-3. Am J Pathol 174:206–15

Rasekh ME, Chiatante G, Miroballo M, Tang J, Ventura M, Amemiya CT, Eichler EE, Antonacci F, Alkan C (2015) Discovery of large genomic inversions using pooled clone sequencing. BioRxiv 015156

Raymond FL, Tarpey P (2006) The genetics of mental retardation. Hum Mol Gen 15(suppl 2):R110-6

Rychlik W, Rhoads RE (1989) A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA. Nucleic Acids Res 17:8543–51

Salm MP, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, Schadt EE, Cookson WO, Wierzbicki AS, Naoumova RP, Shoulders CC (2012) The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Res 22:1144–53

Schilter KF, Reis LM, Schneider A, Bardakjian TM, Abdul-Rahman O, Kozel BA, Zimmerman HH, Broeckel U, Semina EV (2013) Whole-genome copy number variation analysis in anophthalmia and microphthalmia. Clin Genet 84:473–481

Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente RU, Pertz LM, Clark RA, Schwartz S, Segraves R, Oseroff VV, Albertson DG, Pinkel D, Eichler EE (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88

Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135–45

Sie L, Loong S, Tan EK (2009) Utility of lymphoblastoid cell lines. J Neurosci Res 87:1953–9

Stein JM, Bergman W, Fang Y, Davison L, Brensinger C, Robinson MB, Hecht NB, Abel T (2006) Behavioral and neurochemical alterations in mice lacking the RNA-binding protein translin. J Neurosci 26:2184–96

Suktitipat B, Naktang C, Mhuantong W, Tularak T, Artiwet P, Pasomsap E, Jongjaroenprasert W, Fuchareon S, Mahasirimongkol S, Chantratita W, Yimwadsana B (2014) Copy number variation in Thai population. PloS one 9(8):e104355

Tabet AC, Verloes A, Pilorge M, Delaby E, Delorme R, Nygren G, Devillard F, Gérard M, Passemard S, Héron D, Siffroi JP (2015) Complex nature of apparently balanced chromosomal rearrangements in patients with autism spectrum disorder. Mol Autism 6(1):19

Talkowski ME, Ernst C, Heilbut A, Chiang C, Hanscom C, Lindgren A, Kirby A, Liu S, Muddukrishna B, Ohsumi TK, Shen Y, Borowsky M, Daly MJ, Morton CC, Gusella JF (2011) Next-generation sequencing strategies enable routine detection of balanced chromosome rearrangements for clinical diagnostics and genetic research. Am J Hum Genet 88:469–81

The Centre for Applied Genomics Database of Genomic Variants. http://dgv.tcag.ca/dgv/app/home.

Uddin M, Thiruvahindrapuram B, Walker S, Wang Z, Hu P, Lamoureux S, Wei J, MacDonald JR, Pellecchia G, Lu C, Lionel AC, Gazzellone MJ, McLaughlin JR, Brown C, Andrulis IL, Knight JA, Herbrick JA, Wintle RF, Ray P, Stavropoulos DJ, Marshall CR, Scherer SW (2015) A high-resolution copy-number variation resource for clinical and population genetics. Genet Med 17:747–52

Utami KH, Hillmer AM, Aksoy I, Chew EG, Teo AS, Zhang Z, Lee CW, Chen PJ, Seng CC, Ariyaratne PN, Rouam SL, Soo LS, Yousoof S, Prokudin I, Peters G, Collins F, Wilson M, Kakakios A, Haddad G, Menuet A, Perche O, Tay SK, Sung KW, Ruan X, Ruan Y, Liu ET, Briault S, Jamieson RV, Davila S, Cacheux V (2014) Detection of chromosomal breakpoints in patients with developmental delay and speech disorders. PLoS One 9:e90852

van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci U S A 87:1663–7

van Gent DC, Hoeijmakers JH, Kanaar R (2001) Chromosomal stability and the DNA double-stranded break connection. Nat Rev Genet 2:196–206

Vedrenne V, Gowher A, De Lonlay P, Nitschke P, Serre V, Boddaert N, Altuzarra C, Mager-Heckel AM, Chretien F, Entelis N, Munnich A, Tarassov I, Rötig A (2012) Mutation in PNPT1, which encodes a polyribonucleotide nucleotidyltransferase, impairs RNA import into mitochondria and causes respiratory-chain deficiency. Am J Hum Genet 91:912–8

Vergult S, Van Binsbergen E, Sante T, Nowak S, Vanakker O, Claes K, Poppe B, Van der Aa N, Van Roosmalen MJ, Duran K, Tavakoli-Yaraki M (2014) Mate pair sequencing for the detection of chromosomal aberrations in patients with intellectual disability and congenital malformations. Eur J Hum Genet 22(5):652-9

von Ameln S, Wang G, Boulouiz R, Rutherford MA, Smith GM, Li Y, Pogoda HM, Nürnberg G, Stiller B, Volk AE, Borck G, Hong JS, Goodyear RJ, Abidi O, Nürnberg P, Hofmann K, Richardson GP, Hammerschmidt M, Moser T, Wollnik B, Koehler CM, Teitell MA, Barakat A, Kubisch C (2012) A mutation in PNPT1, encoding mitochondrial-RNA-import protein PNPase, causes hereditary hearing loss. Am J Hum Genet 91:919–27

Warburton D (1991) De novo balanced chromosome rearrangements and extra marker chromosomes identified at prenatal diagnosis: clinical significance and distribution of breakpoints. Am J Hum Genet 49:995–1013

Watson JD, Crick FH (1953) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171(4356):737–8

Withers LA, Street HE (1977) Freeze preservation of cultured plant cells. III. The pregrowth phase. Physiologia Plantarum 39(2):171–178

Yakut S, Cetin Z, Sanhal C, Karaman B, Mendilcioglu I, Karauzum SB (2015) Prenatal diagnosis of de novo pericentric inversion inv(2)(p11.2z13). Genetic counseling (Geneva, Switzerland) 26(2):243

Yang L, Lee MS, Lu H, Oh DY, Kim YJ, Park D, Park G, Ren X, Bristow CA, Haseley PS, Lee S (2016) Analyzing Somatic Genome Rearrangements in Human Cancers by Using Whole-Exome Sequencing. Am J Hum Genet 98(5):843–56

Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL (2012) Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. BMC Bioinformatics 13:134

Young I (2005) Medical Genetics. Oxford University Press. New York